AN ABSTRACT OF THE DISSERTATION OF

Noah Fahlgren for the degree of Doctor of Philosophy in Molecular and Cellular Biology presented on December 6, 2010.

Title: Origins and Evolution of Plant *MicroRNA* Genes.

Abstract approved:

James C. Carrington

Eukaryotic small RNA (~20-30 nucleotides) are diverse regulatory molecules that repress gene expression at the transcriptional and post-transcriptional levels, defend hosts against invading viruses and defend genomes against selfish DNA elements. Small RNA populations are studied by high-throughput sequencing of the total small RNA fraction isolated from cells, however, the sequencing depth achieved by next-generation platforms makes genome mapping and analysis computationally intensive with standard methods. Here, methods to generate, parse, map, quantify, standardize and analyze large small RNA data sets are presented. This work demonstrates that small RNA profiling is quantitative and reproducible and that statistical methods can be adapted to facilitate objective comparisons between small RNA and small RNA populations.

Plants RNA silencing systems, including microRNA (miRNA), are important components of complex regulatory networks. Several plant *MIRNA* gene families and their target gene families are ancient, but over two-thirds of *Arabidopsis MIRNA* families are species-specific or restricted to the Brassicaceae lineage. In this work, the repertoires of *MIRNA* in the closely related species *A. thaliana*, *A. lyrata* and *Capsella rubella* were studied. Despite the relatively recent speciation of *A. thaliana* and *A. lyrata* ~10 million years ago, at least 13% of the *MIRNA* from each is species-specific. Additionally, 24–46 *Arabidopsis MIRNA* families arose after the *Arabidopsis*–*Capsella* split ~20 million years ago, supporting a net birth-death rate of 1.2–2.3 *MIRNA* per million years. These data, and data from other species, suggest that *MIRNA* are born and lost frequently throughout the evolution of plants. Further, evidence for the recent origin of 32 *MIRNA* families by duplication events, mostly of protein-coding loci, was demonstrated, but only ~50% of these loci are predicted miRNA targets. Despite the link between *MIRNA* formation and potential target loci, only 25 young *A. thaliana* miRNA have verified targets. As a group, young miRNA tend to be expressed weakly, processed imprecisely and lack biologically relevant targets. Additionally, variation between young *Arabidopsis* miRNA was significantly higher than for ancient miRNA, suggesting that most of the young *MIRNA* are more likely evolving neutrally. Together, the data presented argue that most young *MIRNA* are evolutionarily transient.

Origins and Evolution of Plant *MicroRNA* Genes

by
Noah Fahlgren

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented December 6, 2010
Commencement June 2011

Doctor of Philosophy dissertation of Noah Fahlgren presented on December 6, 2010.

APPROVED:

Major Professor, representing Molecular and Cellular Biology

Director of the Molecular and Cellular Biology Program

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Noah Fahlgren, Author

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to the following people: Jim Carrington for mentoring me during my thesis work; Kristin Kasschau, Alberto Carbonell, Hernan Garcia-Ruiz, Ed Allen, Miya Howell, Atsushi Takeda, Ji Hoon Ahn, Heather Fitzgerald and Zhixin Xie for guidance; Chris Sullivan, Scott Givan and Jason Cumbie for computational guidance, support and discussions; fellow graduate students Elisabeth Chapman, Tai Montgomery, Josh Cuperus, Sanjuro Jogdeo, Seong Jeon Yoo and Steen Hoyer for discussions and comradery; Kathleen Eide, Michelle Yang, Annie Bailey, Amy Shatswell and Heather Sweet for assistance with experiments and plant care; the CGRB office staff Rosa Hill, Gail Millimaki and Goretti Nguyen; Mark Dasenko and Matthew Peterson for assistance with Illumina sequencing; our collaborators from the labs of Detlef Weigel, Nik Grünwald, Todd Mockler, Blake Meyers, Hiro Nonogaki and many others; fellow MCB students; the Beanery and Bombs Away Café for much needed coffee and beer (not at the same time); and friends and family, especially for Kerri. This work was supported by grants awarded to Jim Carrington from the National Science Foundation (MCB-0209836 and MCB-0618433), the National Institutes of Health (AI43288) and the United States Department of Agriculture (2005-35319-15280 and 2006-35301-17420) and the P.F. Yerex and Nellie Buck Yerex Graduate Fellowship awarded to Noah Fahlgren.

CONTRIBUTION OF AUTHORS

Chapter 1: Noah Fahlgren wrote the introduction.

Chapter 2: Jason S. Cumbie, Christopher M. Sullivan, Scott A. Givan and Tyler W. H. Backman wrote and tested the CASHX alignment algorithm and the CASHX pipeline. Kristin D. Kasschau, Elisabeth J. Chapman, Taiowa A. Montgomery, Sunny D. Gilbert and Mark Dasenko generated and sequenced the small RNA libraries. Noah Fahlgren analyzed the sequencing data and did the statistical analyses. Noah Fahlgren, James C. Carrington, Kristin D. Kasschau and Christopher M. Sullivan designed the experiments and wrote the manuscript.

Chapter 3: Theresa F. Law, Sarah R. Grant and Jeffery L. Dangl generated and sequenced small RNA libraries for an analysis included in the original manuscript that is not reproduced here. Noah Fahlgren and Elisabeth J. Chapmen generated the small RNA libraries. Miya D. Howell did the 5' RACE experiments. Kristin D. Kasschau, Christopher M. Sullivan, Jason S. Cumbie and Scott A. Givan helped process the high-throughput sequencing data. Noah Fahlgren identified the novel *MIRNA* genes and did the evolutionary analyses. Noah Fahlgren, James C. Carrington and Kristin D. Kasschau designed the experiments and wrote the manuscript.

Chapter 4: Noah Fahlgren and James C. Carrington designed the experiments. Noah Fahlgren identified orthologous and unique *MIRNA* in three species, analyzed the sequencing data and analyzed the genomic context around each *MIRNA*. Sanjuro Jogdeo analyzed the *MIRNA* nucleotide divergence and miRNA target orthologs. Kristin D. Kasschau, Elisabeth J. Chapman, Sascha Laubinger, Lisa M. Smith and Mark Dasenko generated and sequenced the small RNA libraries. Christopher M. Sullivan and Scott A. Givan provided computational support. Noah Fahlgren, James C. Carrington, Detlef Weigel, Sanjuro Jogdeo, Sascha Laubinger and Lisa M. Smith wrote the manuscript.

Chapter 5: Noah Fahlgren, James C. Carrington and Josh T. Cuperus wrote the manuscript. Additional sections are not reproduced here.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure Page

LIST OF TABLES

General Introduction

Noah Fahlgren

THE ORIGIN OF EUKARYOTIC RNA SILENCING

All forms of life deploy a variety of mechanisms to maintain, replicate and express genetic material. In many cases, these processes are controlled at multiple stages by an assortment of regulatory factors. For example, regulation of gene expression and function can occur at the transcriptional, post-transcriptional or post-translational level by DNA, RNA and protein regulators in cis or trans. In eukaryotes, a large class of short (~20-30 nucleotides) RNA regulators has been described. Eukaryotic small RNA function in diverse pathways, but there are several common themes to all small RNA pathways. A unifying feature of all RNA silencing pathways is that small RNA are loaded into effector complexes containing an ARGONAUTE (AGO) family protein and act as sequence-specific guides that bring AGO proteins in contact with specific target nucleic acids (RNA in most or all cases). AGO-guide-target ternary complexes silence the target either directly (post-transcriptionally) or indirectly by inhibiting transcription of the target locus (Axtell and Bowman, 2008). The form of silencing employed in each pathway is dependent, at least in part, on the AGO effector. Most small RNA are derived from double-stranded RNA by members of the DICER (DCR or DICER-LIKE [DCL]) family (Carthew and Sontheimer, 2009). Functional DCR family proteins generally have two RNase III endonuclease domains that function to cleave small RNA duplexes from longer double-stranded RNA. Many DCR proteins also have PAZ domains that bind free 3' RNA ends and position the RNase III active sites at a fixed molecular distance, making DCR effectively a molecular ruler (Carthew and Sontheimer, 2009). The source of double-stranded RNA is also pathway dependent. Double-stranded RNA can occur naturally through overlapping sense/antisense transcription (in cis or in trans) or through intramolecular base pairing, but can also be generated by RNA-dependent RNA polymerases (RdRP or RDR). Flies and vertebrates lost RdRPs, but they are widely distributed among all other eukaryotes.

Although missing in some lineages, eukaryotic small RNA have been found in five of the six eukaryotic super-groups including opisthokonts (including fungi and animals), Amoebozoa (including slime moulds), Archaeplastida (including land plants and green algae), chromalveolates (including stramenopiles and alveolates) and Excavata, but to date not Rhizaria (Axtell and Bowman, 2008; Roger and Simpson, 2009). Due to the phylogenetic distribution of small RNA and small RNA biogenesis machinery, lineages missing RNA silencing systems are likely derived (Shabalina and Koonin, 2008). For example, the budding yeast *Saccharomyces cerevisiae* lacks RNA silencing, but further analysis of other budding yeast revealed that other *Saccharomyces* species and *Candida albicans* have RNA silencing systems (Drinnenberg et al., 2009). Therefore, evidence suggests that the last common ancestor of all eukaryotes had at least one RNA silencing system that could have involved

DCR, RdRP and AGO (Shabalina and Koonin, 2008). It is difficult to reconstruct the nature of this basal silencing system, but small RNA pathways involved in defense against viruses and transposable genomic elements are more widespread than pathways involved in gene regulation, suggesting that defense may have been the basal function of small RNA-mediated silencing (Axtell and Bowman, 2008; Shabalina and Koonin, 2008). Pathways that regulate endogenous genes may have evolved from this basal defense mechanism through proliferation and adaptation of the core silencing machinery (DCR, AGO and RdRP) and acquisition of accessory factors.

ENDOGENOUS RNA SILENCING PATHWAYS IN PLANTS

MicroRNA

Plant microRNA (miRNA) are predominantly 21-nucleotide small RNA that are processed from longer transcripts. *MIRNA* genes encode non-coding RNAs that are transcribed by RNA Polymerase II and are capped and poly-adenylated (Voinnet, 2009). Most plant *MIRNA* genes are encoded at independent loci, but a few are polycistronic, and some are located in the introns of protein-coding genes and are presumably co-transcribed (Voinnet, 2009). The primary *MIRNA* transcript (pri-miRNA) contains a region of imperfect self-complementarity that folds into a stem-loop RNA helix (Voinnet, 2009). Base pairing between the complementary arms of the stem-loop structure are the source of double-stranded RNA that contains the mature miRNA and miRNA* (small RNA from the opposite arm), therefore no RdRP is required for miRNA biogenesis. Genetic evidence suggests that pri-miRNA are stabilized by the nuclear RNA-binding protein DAWDLE (DDL) (Yu et al., 2008). Stabilized pri-miRNA are routed into nuclear processing centers called D-bodies where they colocalize with the C2H2-zinc finger protein SERRATE (SE), the double-stranded RNA-binding protein HYPONASTIC LEAVES1 (HYL1/DRB1) and DCL1 (Fang and Spector, 2007; Fujioka et al., 2007; Song et al., 2007). SE may interact directly or indirectly with the cap-binding complex (CBC) and is required for efficient miRNA processing, but also has roles in pre-mRNA splicing (Dong et al., 2008; Laubinger et al., 2008). HYL1/DRB1 is also required for accurate miRNA processing, but its functional role is unknown (Kurihara et al., 2006; Dong et al., 2008). SE and HYL1/DRB1 could be involved in the recognition and routing of pri-miRNA and could also be required for the loading or positioning of DCL1. Conversion of pri-miRNA to precursor-miRNA (pre-miRNA) involves DCL1 cleavage of both arms of the stem-loop helix at one end of the *MIRNA*. Most *MIRNA* are processed from the base of the stem-loop first, although the *MIR159/MIR319* family is processed loop-side first (Addo-Quaye et al., 2009; Bologna et al., 2009). Proper positioning of the pri-to-pre-miRNA processing step requires an ~15 base-pair stem and weak or fluid base pairing 1-8 base pairs below the miRNA/miRNA* region (Cuperus et al., 2010a; Mateos et al., 2010; Song et al., 2010; Werner et al., 2010). The RNaseIII domains of DCL1 cleave the *MIRNA* stem at the base of the miRNA/miRNA* region, resulting in a pre-miRNA with a 2-nucleotide 3' single-stranded overhang (Voinnet, 2009). The PAZ domain of DCL1 binds the pre-miRNA overhang and positions the RNaseIII domains ~21 nucleotides away, resulting in the final excision of the miRNA/miRNA* duplex from the pre-miRNA (Voinnet, 2009). The mature miRNA/miRNA* duplex is methylated at both 3' ends by the S-adenosyl methionine-dependent methyltransferase HUA ENHANCER1 (HEN1) (Yu et al., 2005). Methylation is a crucial step in miRNA maturation because 2'-O-methylation stabilizes small RNA by preventing 3' uridylation and degradation by members of the SMALL RNA DEGRADING NUCLEASE (SDN) family (Li et al., 2005; Ramachandran and Chen, 2008). In animals, pre-miRNA are exported to the cytoplasm by exportin-5, and the plant homolog HASTY (HST) is required for accumulation of some mature miRNA, but it is unknown which *MIRNA* substrate are transported through HST (Park et al., 2005; Voinnet, 2009).

The final step in miRNA maturation requires separation of the miRNA and miRNA* strands and incorporation of the mature miRNA into an AGO family protein. Sorting of small RNA into AGO proteins can depend on the 5' nucleotide, the differential thermostability of the miRNA/miRNA* duplex ends and accessory proteins like HYL1/DRB1 (Mi et al., 2008; Montgomery et al., 2008a; Takeda et al., 2008; Eamens et al., 2009). Most *A. thaliana* miRNA associate with AGO1, which prefers miRNA that have 5' uridine residues, but a few associate with AGO2 (5' adenine) or AGO5 (5' cytosine) (Mi et al., 2008; Montgomery et al., 2008a; Cuperus et al., 2010b). Interestingly, AGO7 binds exclusively to miR390 with unknown specificity determinants that do not appear to include the 5' nucleotide (Montgomery et al., 2008a). Mature miRNA act as guides that program AGO proteins to recognize specific target RNAs through base-pairing interactions. Most of the described plant miRNA target sites are highly complementary to the mature miRNA, and in *A. thaliana* only ~1% of transcripts are targeted (~250 genes). These interactions result in target RNA destruction through the slicer activity of AGO1, leading to repression of gene activity (Baumberger and Baulcombe, 2005). However, recent data suggests that plant miRNA-AGO1 complexes can also repress targets by blocking translation initiation or elongation (Brodersen et al., 2008; Lanet et al., 2009). Animal miRNA are predicted to target one-third to one-half of all genes and repress target RNAs primarily without slicing at target sites with limited complementarity (Carthew and Sontheimer, 2009). If non-degradative targeting mechanisms are widespread in plants then the scope of miRNA targeting may be greatly underestimated. Regardless, miRNA-based regulation of many of the sliced targets is critical for proper development and stress response.

For instance, in angiosperms AGO1 is repressed by miR168 and DCL1 is repressed by miR162 and miR838, while in the moss *Physcomitrella patens* DCL1a is repressed by miR1047, demonstrating that miRNA define negative feedback loops on the miRNA pathway itself (Xie et al., 2003; Vaucheret et al., 2004; Rajagopalan et al., 2006; Vaucheret et al., 2006; Axtell et al., 2007). Additionally, nearly half of plant miRNA target transcription factors that are integral components of organ development, developmental timing and hormone signaling regulatory modules (Jones-Rhoades et al., 2006). miRNA are integral parts of more complex regulatory networks and may form several types of circuits by spatially restricting gene expression or creating sharp exclusion boundaries, by dampening or buffering gene expression, by temporal regulation or by altering the morphogenic potential of cells to hormones or other signaling molecules (Rubio-Somoza et al., 2009; Voinnet, 2009).

Short interfering RNA

Small RNA from plant genomes have a bimodal size distribution with peaks of 21 and 24 nucleotides. Mature miRNA represent the bulk of 21-nucleotide small RNA that are found at discrete loci (*MIRNA* genes) throughout the euchromatic arms of each chromosome. In contrast, 24-nucleotide small RNA are more broadly distributed across the genome and are enriched in centromeric and pericentromeric regions. More precisely, the distribution of 24-nucleotide small RNA is highly correlated with the presence of repetitive genomic elements such as transposons and retrotransposons, which are also targets of DNA methylation and heterochromatin formation (Lu et al., 2006; Rajagopalan et al., 2006; Zhang et al., 2006; Kasschau et al., 2007; Zilberman et al., 2007; Lister et al., 2008). Small RNA from these loci are short interfering RNA (siRNA) that are guides that direct the methylation of target loci DNA. How repetitive loci are identified is still poorly understood, but the plant-specific nuclear RNA polymerase IV (Pol IV) is thought to generate single-stranded RNA transcripts from transposon, retrotransposons or other repetitive loci (Pontier et al., 2005; Chan et al., 2006; Haag et al., 2009). Pol IV transcripts are converted to double-stranded RNA by RDR2 and diced into 24-nucleotide siRNA duplexes by DCL3 (Law and Jacobsen, 2010). Mature 24-nucleotide siRNA associate with AGO4-family AGO proteins (in *A. thaliana* AGO4, AGO6 and AGO9) (Qi et al., 2006; Mi et al., 2008; Havecker et al., 2010). Another plant-specific nuclear RNA polymerase, Pol V, generates long non-coding transcripts at repetitive loci and intergenic regions (Wierzbicki et al., 2008). Nascent Pol V transcripts act as scaffolds that recruit AGO4-siRNA complexes (Wierzbicki et al., 2009). It is hypothesized that successful recognition of Pol V transcripts by AGO4-siRNA complexes triggers the recruitment of the de novo DNA methyltransferase DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) and subsequent factors that generate repressive chromatin marks (Law and Jacobsen, 2010). This plant-specific cascade is a genome surveillance system that detects and silences mobile elements to prevent or limit transposition.

Trans-acting siRNA

Although miRNA represent a large proportion of the abundant population of 21-nucleotide small RNA, other 21-nucleotide small RNA populations are found in plant genomes. In *A. thaliana*, four families of non-coding RNAs are the sources of moderately abundant 21-nucleotide trans-acting siRNA (tasiRNA) that match both the sense and antisense strands of the transcripts. The mechanism that routes transcripts from tasiRNA-generating loci (*TAS*) into a siRNA-generating pathway differs between families, but in all cases, miRNA-guided cleavage triggers the production of tasiRNA (Allen et al., 2005). After cleavage, one fragment of the *TAS* transcript is stabilized by SUPPRESSOR OF GENE SILENCING3 (SGS3) and converted into double-stranded RNA by RDR6 (Yoshikawa et al., 2005). DCL4 is recruited to the miRNA-guided cleavage site and sequentially processes the double-stranded RNA into 21-nucleotide tasiRNA duplexes, resulting in the accumulation of tasiRNA in a 21-nucleotide register with the cleavage site (Chen, 2010). The tasiRNA duplex is unwound and one strand is incorporated into an AGO protein. Like miRNA, tasiRNA guide AGO proteins to specific target RNAs in trans based on sequence complementarity. In *A. thaliana*, all of the tasiRNA with known targets have 5’ uridine residues and associate with AGO1 (Mi et al., 2008; Montgomery et al., 2008b).

The *TAS3* family is ancient with members found in the bryophyte moss *P. patens*, the gymnosperm *Pinus taeda* and numerous angiosperms (Axtell et al., 2006). Formation of *TAS3* tasiRNA is initiated by miR390-guided cleavage at a site on the 3’ end of the transcript, however cleavage alone is not sufficient to initiate tasiRNA biogenesis (Axtell et al., 2006; Montgomery et al., 2008a). A second target site on the 5’ end of *TAS3* interacts with AGO7-miR390 complexes, but contains central bulges that prevent transcript cleavage (Axtell et al., 2006; Montgomery et al., 2008a). The function of AGO7-miR390 binding at the 5’ site is unknown, but AGO7 may route *TAS3* transcripts into the tasiRNA-generating pathway or recruit these factors directly. Most *TAS3* transcripts generate two related tasiRNA that target *AUXIN RESPONSE FACTOR* (*ARF*) transcripts. In *A. thaliana*, *TAS3* tasiRNA are part of a leaf patterning and development network where they form a gradient that restricts expression of *ARF3* and *ARF4* to the abaxial leaf cells (Chitwood et al., 2009).

The *TAS1*, *TAS2* and *TAS4* families have only been described in *A. thaliana* and close relatives. Formation of tasiRNA is initiated by miR173- or miR828-guided cleavage of

TAS1/TAS2 and *TAS4*, respectively, but unlike *TAS3*, these are the only miRNA targeting events on these *TAS* transcripts. Additionally, AGO7 is not required for *TAS1*, *TAS2* or *TAS4* tasiRNA formation, and AGO7 does not interact with miR173 or miR828 (Montgomery et al., 2008a). Instead, miR173 and miR828 associate with AGO1, which seems to be sufficient for triggering tasiRNA formation from *TAS1*, *TAS2* and *TAS4* (Mi et al., 2008; Montgomery et al., 2008a; Montgomery et al., 2008b; Chen et al., 2010; Cuperus et al., 2010b). Mature miR173 and miR828 are somewhat unusual in that they accumulate as 22-nucleotide instead of 21-nucleotide species. Recently, it was demonstrated that 22-nucleotide, but not 21-nucleotide, miRNA-AGO1 complexes can trigger RDR6/DCL4-dependent 21-nucleotide siRNA formation from targeted transcripts (Chen et al., 2010; Cuperus et al., 2010b). Therefore, routing of *TAS1*, *TAS2* and *TAS4* transcripts into the tasiRNA pathway is distinct from the *TAS3* pathway. Although *TAS1*, *TAS2* and *TAS4* are young, a similar system has been described in *Oryza sativa* for targets of 22-nucleotide miR2118 and miR2775 (Johnson et al., 2009). Additionally, transcripts from several protein-coding loci are targeted by other 22-nucleotide *A. thaliana* miRNA and generate 21-nucleotide secondary siRNA, suggesting that this phenomenon is not specific to the tasiRNA pathway and may be used to enhance silencing of some targets (Chen et al., 2010; Cuperus et al., 2010b).

DISCUSSION

Plants use diversified RNA silencing systems to regulate the expression of endogenous genes and maintain genome integrity by limiting the spread of mobile elements. Additionally, many of the factors involved in these endogenous systems converge to defend local and systemic tissues against infection by viruses (Ding and Voinnet, 2007). Homologous and analogous systems are found across the eukaryotic domain of life, suggesting that the common ancestor of all extant eukaryotic lineages used small RNA-based regulation. The miRNA pathway is an example of a system that is present in several eukaryotic lineages, although no *MIRNA* genes themselves are conserved between lineages (Axtell and Bowman, 2008; Shabalina and Koonin, 2008). This thesis focuses on the identification and quantitative analysis of miRNA in *A. thaliana* and closely related species to address several important questions regarding the nature of plant miRNA and the evolution of *MIRNA* loci: What does the full repertoire of *MIRNA* genes look like in plants, and how are they quantitatively measured? Where do new *MIRNA* genes originate during genome evolution? How do *MIRNA* genes co-evolve with their targets?

Computational and Analytical Framework for Small RNA Profiling by High-throughput Sequencing

Noah Fahlgren, Christopher M. Sullivan, Kristin D. Kasschau, Elisabeth J. Chapman, Jason S. Cumbie, Taiowa A. Montgomery, Sunny D. Gilbert, Mark Dasenko, Tyler W. H. Backman, Scott A. Givan, and James C. Carrington

RNA, Cold Spring Harbor Laboratory Press
RNA Publication Office,
Center for RNA Molecular Biology,
School of Medicine, Room W127
Case Western Reserve University
10900 Euclid Avenue
Cleveland, Ohio 44106-4973
Volume 15, pages 992-1002
Reprint permission requested

SUMMARY

The advent of high-throughput sequencing (HTS) methods has enabled direct approaches to quantitatively profile small RNA populations. However, these methods have been limited by several factors, including representational artifacts and lack of established statistical methods of analysis. Furthermore, massive HTS data sets present new problems related to data processing and mapping to a reference genome. Here, we show that cluster-based sequencing-by-synthesis technology is highly reproducible as a quantitative profiling tool for several classes of small RNA from *Arabidopsis thaliana*. We introduce the use of synthetic RNA oligoribonucleotide standards to facilitate objective normalization between HTS data sets, and adapt microarray-type methods for statistical analysis of multiple samples. These methods were tested successfully using mutants with small RNA biogenesis (miRNA-defective *dcl1* mutant and siRNA-defective *dcl2 dcl3 dcl4* triple mutant) or effector protein (*ago1* mutant) deficiencies. Computational methods were also developed to rapidly and accurately parse, quantify, and map small RNA data.

INTRODUCTION

Most eukaryotic organisms contain one or more classes of small RNA that function as guides in association with ARGONAUTE (AGO) proteins for regulation at the post-transcriptional or transcriptional level. Diverse small RNA can derive through distinct biogenesis routes and function through specialized classes of AGO proteins (Chapman and Carrington, 2007; Faehnle and Joshua-Tor, 2007; Peters and Meister, 2007; Farazi et al., 2008). microRNA (miRNA) form through multistep processing of self-complementary foldback structures through the activities of DICER (or DICER-LIKE) and other RNaseIII-type nucleases, resulting in products with a 5’ monophosphate and 3’ hydroxyl. Endogenous classes of 5’ monophosphate-containing short interfering RNA (siRNA) form by DICER-mediated processing of long dsRNA, which can arise from bidirectional transcription, self-complementary foldback structures within transcripts, or the activity of RNA-dependent RNA polymerases (RdRPs) (for review, see Voinnet, 2008). Several DICER-dependent classes of siRNA have been characterized in plants (Chapman and Carrington, 2007), and were recently discovered in flies and mice (Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008; Okamura et al., 2008; Tam et al., 2008; Watanabe et al., 2008). In *Caenorhabditis elegans*, secondary siRNA that contain 5’ tri-phosphate arise through an RdRP-dependent, but DICER-independent, route (Pak and Fire, 2007; Sijen et al., 2007). miRNA and siRNA associate with members of the AGO subclass of ARGONAUTE proteins. In animal lineages, several types of small RNA associate with members of the PIWI subclass of ARGONAUTE proteins. These are generally referred to as Piwi-interacting RNA (piRNA), but subtypes include repeat-associated siRNA (flies) and 21U-RNA (*C. elegans*) (Vagin et al., 2006; Brennecke et al., 2007; Batista et al., 2008; Wang and Reinke, 2008). piRNA form through AGO/PIWI-dependent, but DICER-independent, mechanisms, and function to promote germline development and suppress transposons (for review, see Klattenhoff and Theurkauf, 2008). In part, discovery of the existence or expanse of these and other distinct small RNA classes involved high-throughput sequencing, as originally shown by Lu et al., (2005a).

High-throughput sequencing has also emerged as a direct small RNA profiling method (for examples, see Lu et al., 2006; Ruby et al., 2006; Kasschau et al., 2007; Lister et al., 2008). Compared to finite-sample platforms, such as microarray or PCR-based assays, HTS profiling permits semi-open-ended analysis of both known and unknown small RNAs. In principle, because HTS-based profiling is essentially a random-sampling method, the effective linear range should be broad. Picoliter-scale pyrosequencing (Margulies et al., 2005) has been useful for quantitative analysis of small RNA populations in silencing mutants and differential tissues, and in association with specific AGO proteins (Girard et al., 2006; Henderson et al., 2006; Lu et al., 2006; Qi et al., 2006; Rajagopalan et al., 2006; Ruby et al., 2006; Brennecke et al., 2007; Kasschau et al., 2007; Kawamura et al., 2008). This method can be done in a single-sample or multiplexed format. Recently, short-read sequencing-by-synthesis (SBS) of amplified DNA colonies (Bentley, 2006) has emerged as a small RNA profiling method (Czech et al., 2008; Gregory et al., 2008; Lister et al., 2008; Mi et al., 2008; Montgomery et al., 2008a; Okamura et al., 2008; Tam et al., 2008). SBS methods facilitate far greater sequencing depth per sample relative to previous methods. These have been combined with relatively simple analytical methods, usually with relatively low statistical power, for over- or underrepresentation analyses between samples (for examples, see Henderson et al., 2006; Kasschau et al., 2007; Czech et al., 2008; Montgomery et al., 2008a). A consensus regarding rigorous, standardized experimental designs and statistical methods based on replicate samples to analyze individual small RNA or small RNA classes has not yet emerged.

As HTS small RNA data sets increase in size, computational problems intensify. For example, rapid and accurate mapping of small RNA sequences from 10^7 or more reads to a reference genome is a significant computational challenge. Analysis of small RNA data sets presents another set of issues, such as how to normalize data and quantitatively assess differences between multiple samples. Representational artifacts can occur, particularly when the abundance of whole small RNA classes differs significantly between samples. We developed and tested new computational approaches to rapidly parse, quantify, map, and analyze small RNA reads from SBS data sets. We also developed a method using synthetic standards to objectively and quantitatively compare small RNA populations between samples. In addition, we adapted and tested microarray-based statistical methods to identify differentially expressed small RNA between sample sets.

RESULTS AND DISCUSSION

Mapping of SBS reads using CASHX

The SBS platform (Illumina 1G Genome Analyzer) uses bridge-PCR with primers fixed to a silicon slide to amplify DNA clones from single initial molecules (Bentley, 2006). For all small RNA populations analyzed here, an adaptor was ligated to the 3' end in an ATP-independent manner (Pfeffer et al., 2005), and then joined to 5' adaptors by a second enzymatic ligation (Figure S2.1A). cDNA was generated and amplified by 14 PCR cycles, and the resulting population of DNA molecules was subjected to SBS using single-plex loading of individual flow-cell lanes. This generally yielded 5,000,000–9,000,000 raw reads/lane. Raw reads were processed through a pipeline to parse small RNA sequences from the 3' adapter, collapse the data to a uniread set, count the number of reads per unique sequence, map sequences to the reference genome, and annotate sequences with basic information (Figure S2.1B).

Among the steps in the pipeline, accurate mapping of unireads to the reference genome was the most computationally intensive. The traditional DNA search algorithms BLAST and BLAT (Altschul et al., 1990; Kent, 2002) were used in initial tests of randomly sampled populations of 10–10^7 small RNA reads (50% perfect *Arabidopsis* genome match, 50% mismatch). At 10^4 queries and higher, BLAT performed faster than BLAST (Figure 2.1). For example, BLAT mapped 10^6 reads 3.2-fold faster than BLAST (Figure 2.1). The faster speed of BLAT with larger read sets is due to the database indexing method (Kent, 2002). However, at 10^7 reads, BLAT required ~78.8 h, which was judged to be unacceptably slow for SBS data sets.

An alternative mapping program, cache-assisted hash search with XOR digital logic (CASHX), was developed to map small RNA reads efficiently to a reference genome. This program utilizes a 2 bit-per-base binary format of query and reference genome sequences to reduce computational weight. The reference genome is divided into all possible 30 nucleotide (nt) sequences, each of which is linked to data for chromosome, strand, and start/end coordinates. Each 30-mer is indexed by a preamble string of 4 nucleotides at the 5' end within a HASH database. The initial HASH database, therefore, has 256 (4^4) containers of 30-mer sequences, where each sequence within a container has the same first four nucleotides. The CASHX algorithm searches the HASH index in 0(1) constant time (fast) and the containers in 0(1) linear time (slow). Therefore, the amount of data within a container impacts processing speed disproportionately compared to the number of indexed containers. To increase processing speed, the HASH database, indexed to a 4-nucleotide preamble, is easily transformed to a user-defined preamble string of 8–12 nucleotides to optimize the number of containers with the number of sequences in each container. In the case of a 12-nucleotide preamble, the CASHX database built from the *Arabidopsis* genome was created in less than 8 min, used 7.2G of memory, and generated 16,777,216 containers of 30-mer sequences.

Next, the genome HASH database is searched with each small RNA-derived query sequence. First, the query preamble sequence is identified within the HASH database using key value pairs, thereby locating a container. This search can be done after preloading the HASH database into cache memory, or by searching directly from file space. If the HASH database is not precached, a key value pair hit loads the container contents into memory. Second, each sequence within a hit container is searched using an XOR digital logic string. Sequences that pass through the XOR gate with an outcome of zero correspond to a perfect match. Default CASHX output files contain sequence information, number of reads/sequence in the library, and a list of perfect genome hits, including strand and start/stop coordinates. The output can also be formatted for compatibility with BLAT PSL/PSLX formats (Kent, 2002). The minimum searchable sequence length is 15 nucleotides. Sequences over 30 nucleotides in length are divided into 30-mers and aligned to the CASHX HASH database. Consecutive hits on the genome are identified to reconstruct the full sequence match. CASHX was tested successfully using sequences up to 10,000 nucleotides in length.

CASHX was tested using 10–10^8 sequences (50% *Arabidopsis* genome matched, 50% mismatched), with and without precaching of the HASH database. Without precaching, processing time for 10^3 queries was comparable to BLAT and BLAST (Figure 2.1). However, CASHX processing speed accelerated as numbers of queries increased above 10^3. This was due to the impact of on-the-fly data caching of recurring searches within a given container, and because searching in cache memory space is significantly faster than searching in file space. For example, 10^3 CASHX searches done after precaching finished ~500-fold faster than the same number of CASHX searches done using file space (Figure 2.1). Compared to BLAT, CASHX run with precaching was ~500–900-fold faster for 10^3 or more queries (Figure 2.1). Only CASHX performed at speeds deemed practical under normal circumstances with 10^7 queries or greater.

Other programs, such as ELAND (Illumina, http://www. illumina.com) and SOAP (Li et al., 2008), can be used to map HTS reads to a reference genome. Using a 5’ ligation-dependent SBS data set of *Arabidopsis* small RNA (6,668,228 parsed reads of 18–29 nucleotides), ELAND and SOAP both identified reads with genomic hits with speed comparable to, or slightly slower than, CASHX (Table 2.1). All reads and unique sequences returned using CASHX were returned with ELAND, and these were confirmed to be bona fide hits to the *Arabidopsis* genome by using a direct string comparison between the query sequence and the sequence retrieved by FASTACMD (Johnson et al., 2008) using the coordinates supplied by CASHX. SOAP, which was run using an 8-nucleotide seed size and searching for reads with zero mismatches, returned fewer total reads and unique sequences. CASHX identified more genomic loci with perfect hits than did ELAND or SOAP (Table 2.1). ELAND returns fewer genomic hits because repetitive hits are reported only once. The basis for fewer hits identified by SOAP is not clear, but may relate to the method of reference genome indexing used prior to search.

Reproducibility of SBS small RNA data sets

To assess the SBS method as a quantitative profiling tool, biological and technical replicates of *Arabidopsis* small RNA SBS runs were compared for reproducibility. Two small RNA classes—miRNA families, and 24-nucleotide siRNA within genomic windows or bins (50,000 nucleotides, 10,000 nucleotides scroll)— were quantified separately. Read counts were normalized to adjust for differences in library size, or sequencing depth, to reads per million (RPM). Data were also repeat normalized to distribute read counts evenly among all loci with a perfect match, and small RNA corresponding to rRNA, tRNA, snRNA, and snoRNA were removed.

The correlation of normalized miRNA family and 24-nucleotide siRNA bin reads between biological (two distinct samples processed independently and run on two lanes) replicates was very high (Spearman's ρ for miRNA and 24-nucleotide siRNA bins was 0.947261 and 0.9453777, respectively) (Figure 2.2A, upper panel). For technical replicates (one amplicon preparation divided among two lanes in one flow cell), the correlation for normalized miRNA family and 24-nucleotide siRNA bin reads was even higher (Spearman's ρ = 0.9818703 and 0.9569397, respectively) (Figure 2.2B, upper panel).

Despite the high degree of correlation between replicates, variability of individual miRNA family or 24-nucleotide siRNA bins was not constant. Residual error after fitting a linear model was low for highly abundant small RNA, while residual error for less abundant small RNA was high (Figure 2.2A,B, lower panels). This nonuniform variance, or heteroscedasticity, was significant ($P \leq 4.9 \times 10^{-4}$, Goldfeld–Quandt and Breusch–Pagan tests). Heteroscedasticity is also a common feature of microarray data where probes associated with higher intensity signal are more reproducibly measured than probes with lower signal (Fan et al., 2004). Significant heteroscedasticity will cause standard error estimates to be inflated, resulting in a decrease in statistical power and an increase in Type I error (Allison, 2006; Montgomery et al., 2006). Various techniques, such as those described by Tusher et al., (2001), attempt to correct for this expression level-dependent variance.

Spike-in standards to compare small RNA SBS data sets

Although the trends shown in Figure 2.2 are clear, the profiling methods presented above suffer from a lack of objective standards against which to compare small RNA data. Thus, given the semi-open-ended nature of the SBS platform, comparisons between different small RNA depend largely on relative, not absolute, abundance. To overcome this limitation, three unique oligoribonucleotides (Std2, Std3, Std6) were designed to mimic canonical 21-nucleotide small RNA (5’ monophosphate, 3’ hydroxyl) and tested as spike-in standards with small RNA SBS libraries (Figure 2.3A). In initial tests, the three oligoribonucleotides were each added to four *Arabidopsis* total RNA samples (100 µg) in four amounts (0.01, 0.1, 1.0, and 10.0 pmol), and preparations were subjected to SBS sequencing. Note that the standards were included in all preparatory steps, including the initial purification of small RNA by gel elution. In each library, the normalized reads for each of Std2, Std3, and Std6 were similar and read counts increased in a linear progression in samples containing between 0.01 and 1.0 pmol initial spike-in amounts (Figure 2.3A). At 10 pmol, however, standard reads approached saturation on the flow cell (Figure 2.3A). Additionally, five other oligoribonucleotide standards were tested in the same concentration range, each yielding linear increases between 0.01 and 1.0 pmol initial amounts, although the efficiency of sequencing individual RNA varied among different standards (data not shown).

Given the similar efficiencies of sequencing Std2, Std3, and Std6, a spike-in cocktail containing standards in three different amounts (Std 2, 0.01; Std 3, 0.1; Std 6, 1.0 pmol) per 100 µg total sample RNA was tested in two small RNA profiling experiments. In one, small RNA were compared between wild type *Arabidopsis* (Col-0) and the *dcl1-7* mutant, which accumulates lower levels of most miRNA (Park et al., 2002; Reinhart et al., 2002). In the other, wild type was compared to the *dcl2-1 dcl3-1 dcl4-2* triple mutant, which is deficient in all known classes of siRNA (Zhang, 2008). The standards formed an objective reference curve against which experimental samples, including canonical miRNA, canonical miRNA*, 21-nucleotide tasiRNA, and 24-nucleotide siRNA bins (1000-nucleotide windows, 200-nucleotide scroll), were normalized and converted to picomoles.

In the *dcl2-1 dcl3-1 dcl4-2* triple mutant, tasiRNA and 24-nucleotide siRNA populations were depressed relative to the standard curve, while miRNA and miRNA* were generally unaffected (Figure 2.3B, left). Conversely, in the *dcl1-7* sample, miRNA, miRNA*, and tasiRNA reads were generally low relative to the standard curve, whereas 24-nucleotide siRNA were largely unaffected (Figure 2.3B, right). Loss of tasiRNA in the *dcl1-7* mutant was expected, as tasiRNA biogenesis requires transcript cleavage at a miRNA target site. These differences were analyzed statistically using a nonparametric *t*-test (Mann–Whitney–U/Wilcoxon rank sum test) to compare the quantities of miRNA and miRNA*, tasiRNA, and 24-nucleotide siRNA bins between wild type and each mutant sample. As expected, miRNA, miRNA*, and tasiRNA groups were significantly underrepresented in *dcl1-7* versus wild type (miRNA and miRNA* were 5.5-fold underrepresented, $P = 4.04 \times 10^{-8}$; tasiRNA were ~32-fold underrepresented, $P < 2.2 \times 10^{-16}$). The 24-nucleotide siRNA class was not significantly affected in the *dcl1-7* mutant ($P = 0.1943$). In contrast, miRNA and miRNA* were not significantly affected in the *dcl2-1 dcl3-1 dcl4-2* triple mutant ($P = 0.311$), but tasiRNA and 24-nucleotide siRNA were significantly underrepresented (tasiRNA were ~33-fold underrepresented, $P < 2.2 \times 10^{-16}$; 24-nucleotide siRNA were ~18-fold underrepresented, $P < 2.2 \times 10^{-16}$).

These data indicate that synthetic oligoribonucleotide standards can be used effectively in profiling experiments to objectively compare and normalize small RNA populations between independent samples.

Statistical analysis of small RNA profiles

Although comparisons between entire small RNA populations are often useful, identification of differentially expressed individual small RNA is often desired. Meaningful analysis of small RNA differences between samples requires statistical power at rigorous significance levels, and this is obtained through increased sample size (replicate data sets). We reasoned that abundance of a given sequence in an SBS data set is roughly analogous to signal intensity in a single channel microarray experiment. The commonly used significance analysis of microarrays (SAM) method (Tusher et al., 2001) was adapted. Library size-normalized small RNA from immature flowers of wild type Col-0 and *ago1-25* mutant plants, each of which was represented by three biological replicates (Montgomery et al., 2008b), were compared. miRNA and tasiRNA, but not 24-nucleotide siRNA, are known to be moderately affected in the hypomorphic *ago1-25* mutant (Morel et al., 2002). Therefore, we considered these data sets to be well suited for testing the statistical power and sensitivity of SAM as applied to sequencing-by-synthesis data (SAM-seq). The SAM procedure uses a relative difference score $d(i)$ as a statistic to test for significant differential expression (Tusher et al., 2001). The $d(i)$ are compared to a null distribution of scores, determined using a permutation-based resampling method, to determine the significance of individual scores. In this case, the relative difference of a small RNA between *ago1-25* and wild type Col-0 is

$$d(i) = \frac{\bar{x}_t(i) - \bar{x}_u(i)}{s(i) + s_0},$$

where $\bar{x}_t(i)$ and $\bar{x}_u(i)$ are the average RPM for the *i*th small RNA in *ago1-25* and wild type Col-0, respectively. The value *s*(*i*) is the combined standard deviation of replicate measurements of small RNA *i* in both *ago1-25* and wild type Col-0 and is given by

$$s(i) = \sqrt{\left(\frac{1/n + 1/m}{n + m - 2}\right)\left(\sum_t [x_t(i) - \bar{x}_t(i)]^2 + \sum_u [x_u(i) - \bar{x}_u(i)]^2\right)},$$

where *n* and *m* are the number of replicates in *ago1-25* and Col-0, respectively. The value s_0 is a small, positive constant that is used to minimize the coefficient of variation for the data set (Tusher et al., 2001). Therefore, *d*(*i*) is essentially a *t*-statistic that is modified to use a variance shrinkage procedure that increases inferential power (for review, see Allison et al., 2006). Also, rather than using a *P*-value cutoff, SAM allows for control of the false discovery rate (FDR)—the percentage of false positive tests expected out of all tests called significant—using a *Q*-value measure (Tusher et al., 2001). In an initial series of tests, a post hoc power analysis, with varying sample size (two, three, four, or five replicates), was done using the SAM package for R (Tibshirani, 2006; R Development Core Team, 2009). The power of a test (power = 1-false negative rate [FNR]) is affected by both sample size and effect size (Tibshirani, 2006). For instance, with duplicate samples for wild type Col-0 and *ago1-25* and a small effect size (100 small RNA) the FNR was estimated to be over 90% (Figure 2.4A). Although the FNR decreases with increasing effect size, addition of replicates also decreased the FNR (Figure 2.4A). For instance, with three replicates compared to two replicates, the FNR was decreased by almost 40%; however, even with five replicates, the FNR is still predicted to range from ~5%–25%, depending on effect size (Figure 2.4A). In contrast, the FDR, which is controlled for using SAM-seq, was predicted to be relatively low (generally under 5%) for all replicated sample sizes and effect sizes (Figure 2.4A). Therefore, increased numbers of replicate data sets are needed to increase statistical sensitivity (decrease the FNR).

Next, unique *Arabidopsis* small RNA sequences (n = 359,447) that were identified in at least three samples and did not originate from rRNA, tRNA, snRNA, or snoRNA were analyzed. SAM-seq yielded 1161 differentially expressed small RNA (321 over- and 840 underrepresented in *ago1-25*) that were each >twofold affected, with a FDR < 0.041 (Figures 2.4B–D, 2.5; see Supplemental Table 1 online; http://rnajournal.cshlp.org). Of course, not all small RNA that were twofold-affected, particularly those with low read numbers, were called significant (Figures 2.4C,D, 2.5). SAM-seq predicted that the FNR for underrepresented small RNA with *d*(*i*) between -0.13 and -0.75 was greater than 7%, while the FNR for overrepresented small RNA was 0% (Table S2.1). This indicates that many more small RNA may be underrepresented in the *ago1-25* mutant than can be identified with the statistical power available with three replicates.

To assess the results from SAM-seq, significantly over- and underrepresented small RNA were categorized and analyzed for 5' nucleotide content. Previous analyses revealed the preference of AGO1 for miRNA and tasiRNA with a 5' uracil (5'U) (Mi et al., 2008; Montgomery et al., 2008a); therefore, 5'U-containing miRNA and tasiRNA were predicted to be disproportionately affected in *ago1-25* plants. Based on annotated features, 58% and 38% of over- and underrepresented small RNA, respectively, were categorized as miRNA, miRNA*, tasiRNA, known phased siRNA or inverted repeat-derived siRNA (Figure 2.6). Most overrepresented small RNA were derived from two large inverted repeats on chromosome 3, or corresponded to inaccurately processed or nonannotated species from *MIRNA* foldbacks (Figure 2.6). Underrepresented small RNA were fairly evenly distributed among miRNA, tasiRNA, a collection of 21-nucleotide siRNA from mRNA encoding pentatricopeptide repeat (PPR) proteins (Howell et al., 2007), and the inverted repeats from chromosome 3 (Figure 2.6). However, underrepresented small RNA overwhelmingly possessed a 5'U, while overrepresented small RNA most often possessed a 5' base other than a 5'U ($P < 2.2 \times 10^{-16}$ and $P = 4.4 \times 10^{-4}$, respectively, Fisher's exact test) (Figure 2.6).

Interestingly, three new miRNA were identified in the significantly affected sets (Table S2.2). Two miRNA (miR1886.2 and miR2111a/b) possessed a 5'U and were in the underrepresented set. The third miRNA (miR2112), which derives from an intron at the At1g01650 locus, possessed a cytosine base at the 5' end of the major small RNA from both the 5' and 3' arms (Table S2.2). Each of these was represented by miRNA and miRNA* sequences in these or previously published libraries (Rajagopalan et al., 2006) and originated from foldbacks that fulfill consensus *MIRNA* requirements (Figure S2.2; Ambros et al., 2003; Meyers et al., 2008), but none were conserved in poplar, cassava or rice. miR1886.2 was previously identified as a candidate miRNA by Rajagopalan et al., (2006) and is derived from the recently identified *MIR1886* foldback (German et al., 2008). miR1886.2 is the most abundant small RNA from *MIR1886*, according to small RNA libraries available at the *Arabidopsis* SBS database (http://mpss.udel.edu/at_sbs), and is offset from the annotated miR1886.1 sequence by 9 nucleotides (Figure S2.2). Both miR1886.1 and miR1886.2 have miRNA* sequences represented in the public databases (ASRP, http://asrp.cgrb.oregonstate.edu/db and *Arabidopsis* SBS, http://mpss.udel.edu/at_sbs) and may represent abundant, offset variants like those seen from *MIR161* (Allen et al., 2004).

In addition to analysis of individual sequences, SAM-seq was used to identify AGO1-dependent siRNA clusters. Such clusters may be composed of sets of low-abundance siRNA that, individually, may occur at levels insufficient for reliable SAM-seq analysis. Reads in each sample were library size-normalized, repeat-normalized, then masked for previously annotated miRNA, tasiRNA, and sequences from rRNA, tRNA, snRNA and snoRNA. siRNA of 21 or 24 nucleotides were independently binned using the scrolling window method (1000-nucleotide window, 200-nucleotide scroll). Only bins with reads from at least three replicates were included in the analysis. SAM-seq identified 146 differentially represented 21-nucleotide siRNA bins (16 over- and 130 underrepresented in *ago1-25*) that were all more than twofold affected, with a FDR < 0.046 (Figure 2.7A, left; see Supplemental Table 4 online; http://rnajournal.cshlp.org). SAM-seq also identified 276 differentially expressed 24-nucleotide siRNA bins (114 over- and 162 underrepresented in *ago1-25*) that were all more than twofold affected, with a FDR < 0.035 (Figure 2.7B, left; see Supplemental Table 5 online; http://rnajournal.cshlp.org). The majority (68.5%) of down-affected 21-nucleotide siRNA bins corresponded to RDR6/DCL4-dependent siRNA, such as those from the *PPR* gene family (Axtell et al., 2006; Lu et al., 2006; Rajagopalan et al., 2006; Howell et al., 2007), that were identified previously based on different criteria (Figure 2.7A, right). In fact, 90.5% of all *PPR* loci that were identified previously as RDR6/DCL4-dependent siRNA-generating loci were recognized as significantly affected in this analysis (Figure 2.7A, right). The other 21-nucleotide siRNA-generating loci corresponded to 31% of all other (non-*PPR*) such loci identified previously (Figure 2.7A, right). Additionally, two of the underrepresented 21-nucleotide bins represented heterogeneous small RNA derived from the *MIR839* locus, and three corresponded to two of the novel or recently identified miRNA families identified above (*MIR1886* and *MIR2111*) (Table S2.2). Among the remaining over- and underrepresented 21 and 24-nucleotide bins, 124 corresponded to transposable element loci, and 104 did not overlap any currently annotated features (Figure 2.7A,B). The far greater number of underrepresented, compared to overrepresented, 21-nucleotide siRNA bins is likely a reflection of direct or indirect dependence of 21-nucleotide siRNA on AGO1 for stability or biogenesis.

CONCLUSIONS

In this study, we presented methods to generate, parse, map, quantify, standardize, and analyze large SBS-derived data sets, and demonstrated that SBS profiling of diverse small RNA populations can be done quantitatively and reproducibly. Although in this study we generated SBS data for *Arabidopsis* small RNA populations, these methods are not limited to plants. We introduce CASHX, a new mapping program developed to identify and quantify perfect genome hits for HTS data sets to a reference genome. Along with some other search programs (for example, Ning et al., 2001; Kahveci and Singh, 2003), CASHX takes advantage of HASH database structure and cache memory to rapidly identify genome loci with matches to small RNA. Additionally, CASHX is not limited to small RNA data sets. It also works well with other types of SBS data, such as those resulting from mRNA transcript profiling or genomic resequencing (data not shown). Additionally, the CASHX pipeline is suitable for processing SAGE-like reads containing an adaptor linked to a cDNA sequence tag, as the CASHX parsing tool effectively separates the adaptor from tag sequence before alignment to the reference genome. CASHX can also work with longer reads, such as those produced by 454 pyrosequencing (Margulies et al., 2005) or traditional Sanger sequencing.

The application of SBS as a small RNA profiling tool is enabled by the high quantitative reproducibility between like samples using the Illumina platform. Even with the consistency between replicates, we show that the use of synthetic oligoribonucleotides as spike-in standards can facilitate more objective, quantitative comparisons of small RNA data sets from different samples, and should reduce problems associated with interpreting proportional representation differences. Although we did not multiplex samples in this analysis, variant (barcoded) synthetic standards should work equally well with mixed samples.

We also demonstrate the usefulness of adapting the microarray-based method SAM (Tusher et al., 2001) as a statistical method for analyzing replicate SBS data sets. By using SAM-seq, we were able to detect individual, differentially expressed small RNA or differentially expressed small RNA clusters with a low false discovery rate. The false negative rate, or sensitivity, of SAM-seq is limited by number of replicate samples. The applicability of SAM-seq to SBS data sets was shown through quantitative discrimination of known small RNA classes and subclasses in wild type and small RNA-defective mutants. The utility was also demonstrated through discovery of new miRNA based on a quantitative threshold.

The profiling methods presented here are not without shortcomings. Most notably, low-abundance small RNA are difficult to analyze with high statistical confidence. Improvements in SBS or other ultra-high-throughput sequencing technology will undoubtedly lower the abundance threshold at which significant calls can be made. Additionally, the use of spike-in standards as objective references is limited by the relatively few data points compared to the number of experimental data points. Inclusion of additional standards will reduce this limitation.

MATERIALS AND METHODS

Plant materials and small RNA libraries

Mutant lines (Col-0 background) included *dcl1-7* (Xie et al., 2005b), *dcl2-1 dcl3-1 dcl4-2* (Deleris et al., 2006), and *ago1-25* (Morel et al., 2002). Small RNA libraries were constructed as in Kasschau et al., (2007), but with the following modifications. Spike-in control oligoribonucleotides were added to 100 µg of total RNA before amplicon preparation was started (see above). The 3' adaptor was replaced with the miRNA cloning linker-1 (Integrated DNA Technologies, www.idtdna.com), which is 5' adenylated to allow for ATP-independent ligation, and has a 3' dideoxycytosine to prevent adaptor self ligation. The 5' adaptor was replaced with an RNA oligonucleotide (5'-GUUCAGAGUUCUACAGUCCGACGAUC-3'). cDNA was amplified by PCR using Phusion High-Fidelity DNA Polymerase (New England Biolabs, www.neb.com), 5' PCR primer (5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3'), and 3' PCR primer (5'-CAAGCAGAAGACGGCATACGAATTGATGGTGCCTACAG-3'). PCR primers contained sequences required for cluster generation on the Illumina Genome Analyzer system (Illumina, http://www. illumina.com). DNA amplicons were recovered from preparative 6% polyacrylamide gels by electro-transfer to DE81 paper, high-salt elution, and ethanol precipitation. DNA amplicons were then sequenced (36 cycles) using an Illumina 1G (Illumina, http://www. illumina.com). Amplicons (2.5 pmol) were added to each flow-cell lane. Sequencing primer (5'-GTTCAGAGTTCTACAGTCCGA-3'; 200 pmol/mL working stock) was prepared according to the Illumina protocol. A current, full protocol is available at http://jcclab.science.oregonstate.edu/?q=node/view/54596.

CASHX programs and scripts

The CASHX package is available at http://jcclab.science.oregonstate.edu/?q=node/view/54596. The suite is composed of Perl scripts and C++ programs that parse, quantify, and map reads and populate the resulting data into a MySQL database (Figure S2.1B). The mapping component of the CASHX package contains several programs and scripts for CASHX database formatting, database searching, and result processing. The program cashx_formatDB is used to convert the reference sequence from FASTA format to a HASH-indexed, 2-bit-per-base binary format. cashx_formatDB uses file space for work to minimize memory requirements, but as a result, is relatively slow. CASHX databases can also be created with the much faster program cashx_formatDBmem, which uses memory instead of file space (Figure S2.1B). Additional details about CASHX are provided as Supplementary Information.

Statistical analyses

All statistical analyses were done using R v2.7.0 (R Development Core Team, 2009). For comparisons shown in Figure 2.2, an ordinary least squares (OLS) linear model (R function ‘‘lm,’’ ‘‘stats’’ package (R Development Core Team, 2009)) was fitted to the miRNA families and 24-nucleotide siRNA bins. The residual error analysis was done by plotting the absolute value of the residual error for each miRNA family or bin from the linear model versus the fitted y-values. The Goldfeld–Quandt and Breusch–Pagan tests (R functions ‘‘gqtest’’ and ‘‘bptest,’’ ‘‘lmtest’’ package (Zeileis and Hothorn, 2002)) were used to check for significant heteroscedasticity.

For comparisons involving the spike-in standards shown in Figure 2.3, a standard curve was generated for each sample (wild type Col-0, *dcl1-7* mutant and *dcl2-1 dcl3-1 dcl4-2* triple mutant) by fitting an OLS linear model to the log-transformed oligoribonucleotide standard reads versus the log-transformed pmol of standard. The standard curves were of the general form $p = mr+b$, where p is pmol of small RNA, r is reads and m and b are the slope and intercept of the model, respectively. The standard curves were $p = 1.105r - 16.54$ for wild type Col-0, $p = 1.049r - 15.85$ for *dcl1-7* mutant and $p = 1.008r - 15.29$ for *dcl2-1 dcl3-1 dcl4-2* triple mutant. Plugging in log-transformed observed reads for r returns log-transformed pmol, and after back transformation returns an estimate of pmol of a particular small RNA in the original 100 µg of total RNA. Small RNA quantities were calculated for each canonical (previously annotated) miRNA and miRNA*, for each possible 21-nucleotide tasiRNA from any of the eight *TAS* loci, and for 24-nucleotide siRNA bins (1000-nucleotide windows, 200-nucleotide scroll). A Mann–Whitney–U/Wilcoxon rank sum test (R function ‘‘wilcox.test,’’ ‘‘stats’’ package (R Development Core Team, 2009)) was used to evaluate whether there were small RNA population level differences between wild type and either *dcl1-7* mutant or *dcl2-1 dcl3-1 dcl4-2* triple mutant samples.

All SAM-seq analyses were done by adapting the samr package for R (Tusher et al., 2001). Statistical power analysis was done using the samr ‘‘samr.assess.samplesize’’ function for two, three, four, or five replicates per sample and an expected difference of twofold. The SAM-seq analysis was done using the ‘‘samr’’ function with two class unpaired data, a standard test statistic, noncentering, and 1000 permutations. A delta value was chosen that kept the FDR less than 0.05 for each SAM-seq run. Delta tables for all SAM-seq analyses are available in supplemental information (see Supplemental Tables 6-8 online; http://rnajournal.cshlp.org). Miss rate tables for all SAM-seq analyses are also available in supplemental data (Table S2.1, see Supplemental Tables 9 and 10 online; http://rnajournal.cshlp.org).

Small RNA data sets

SBS small RNA data sets used in this paper are available from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo). GEO accessions are as follows: biological and technical replicates of wild type Col-0 (GSE14694) and wild type Col-0, *dcl1-7* mutant and *dcl2-1 dcl3-1 dcl4-2* triple mutant plants (GSE14695). Data for wild type Col-0 and *ago1-25* mutant flower tissue were previously described by Montgomery et al., (2008b) and are available from GEO (GSE13605).

ACKNOWLEDGMENTS

We thank Goretti Nguyen for assistance with small RNA library preparations, and Amy Shatswell for technical assistance. We also thank Detlef Weigel, Stefan Henz, Korbinian Schneeberger, and Stephan Ossowski for helpful discussions. This work was supported by grants from NSF (MCB-0618433), NIH (AI43288), and USDA-NRI (2006-35301-17420) to J.C.C., and a postdoctoral fellowship from the Damon Runyon Cancer Fund to S.D.G.

Table 2.1. Perfect *Arabidopsis* genomic hits identified by three programs

	Reads with genomic hits[a]			
Program	**Total**	**Unique**	**Genomic hits**	**Time**
CASHX[b]	3,808,746	314,626	892,775	00:02:13.20
ELAND	3,808,746	314,626	884,978	00:13:17.54
SOAP	3,518,137	293,548	837,079	00:17:37.74

[a] Data were derived from 6,668,228 total parsed small RNA reads. Sequencing error accounts for the vast majority of reads that fail to match the *Arabidopsis* genome.

[b] Genomic hits verified using FASTACMD.

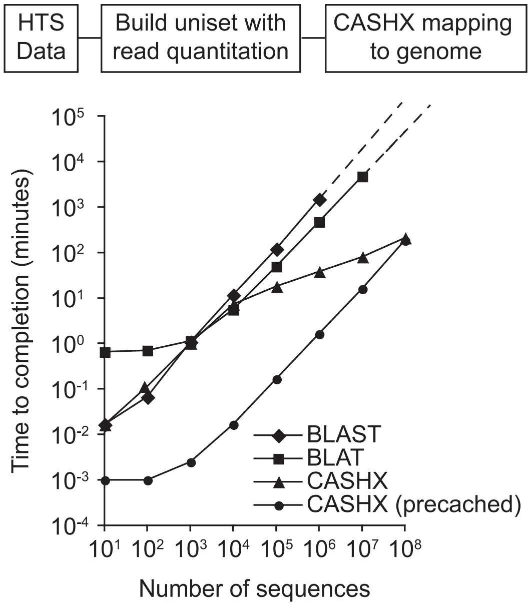


Figure 2.1. Processing speed to query 10–10^8 small RNA sequences (50% *Arabidopsis* genome perfect match, 50% mismatch) using BLAT, BLAST, and CASHX.
Each data point represents the average of five independent runs. CASHX was run with and without precaching. Due to the extensive time requirement, a maximum of 10^6 and 10^7 queries were done by BLAST and BLAT, respectively.

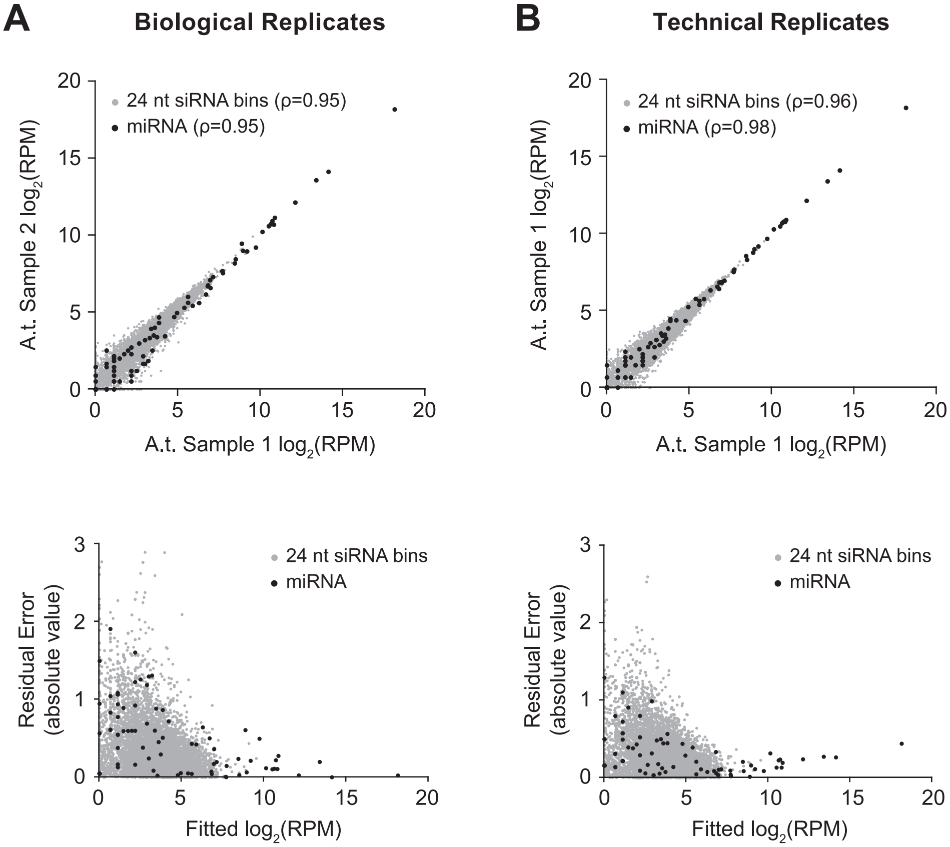


Figure 2.2. Reproducibility of SBS data sets.
Comparison of **(A)** biological and **(B)** technical replicates of Arabidopsis small RNA samples prepared using the 59 ligation-dependent amplicon preparation method. Upper graphs show normalized small RNA reads in one replicate versus another. Lower graphs show absolute residual error versus fitted read values. In all graphs, miRNA are plotted as black dots and 24-nucleotide siRNA bins (50,000-nucleotide windows, 10,000-nucleotide scroll) are plotted as gray dots. Data were normalized to reads/million.

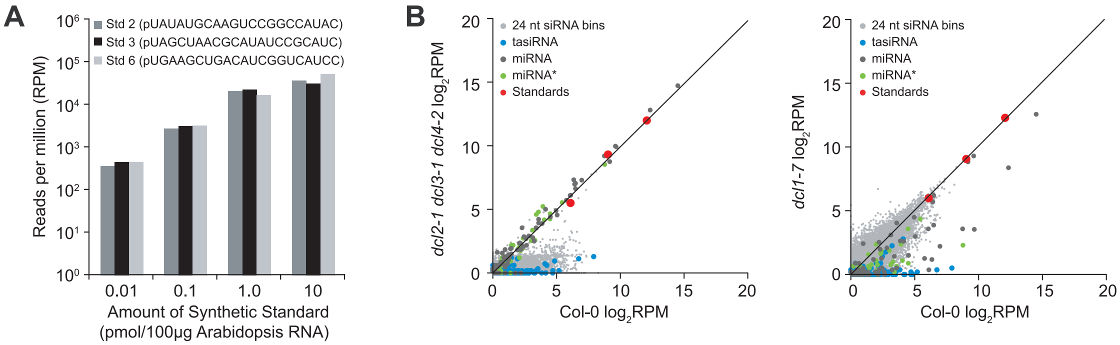


Figure 2.3. Use of synthetic, 21-nucleotide oligoribonucleotide standard spike-in controls for SBS sequencing.

(A) Comparison of normalized reads for Std2, Std3, and Std6 added to *Arabidopsis* total RNA samples at four different amounts. Data show reads per million (RPM). **(B)** Comparison of standards, miRNA, miRNA*, tasiRNA, and 24-nucleotide siRNA bins in samples prepared from wild type (Col-0) and *dcl2-1 dcl3-1 dcl4-2* (left) or *dcl1-7* (right) mutant plants. Scatter plots show log2-scale RPM for each small RNA class or standard. Standards are shown with a fitted linear regression model.

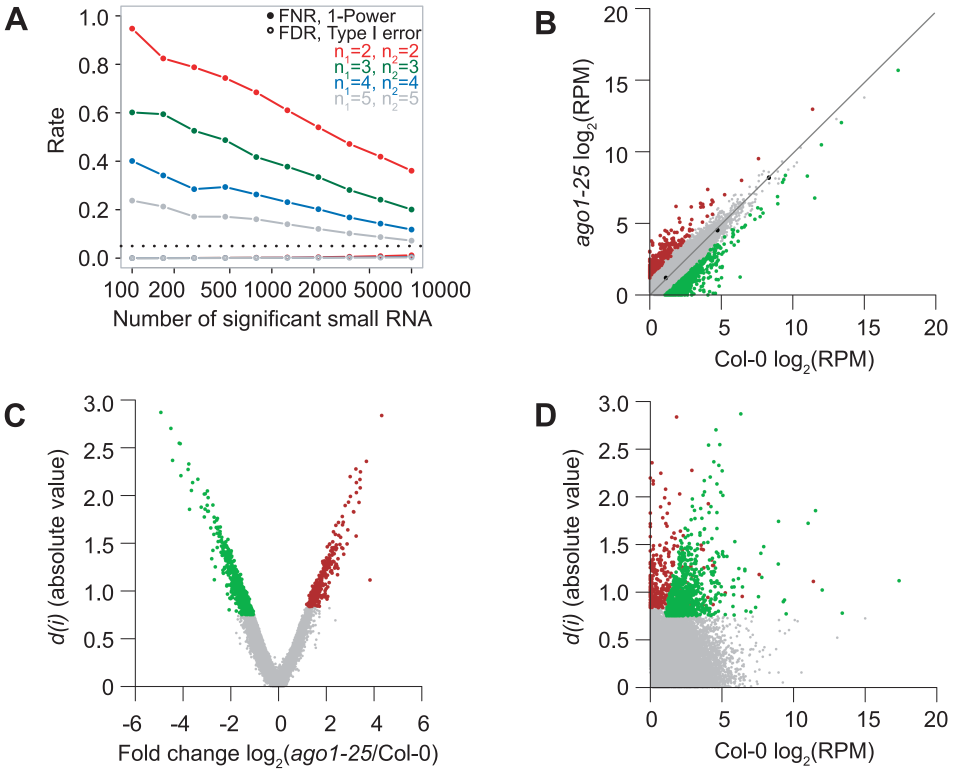


Figure 2.4. SAM-seq analysis of differentially expressed small RNA.
(A) Statistical power assessment for the Col-0 versus *ago1-25* analysis with 2, 3, 4, or 5 replicates modeled per sample. **(B)** Scatter plot of the mean (n = 3) RPM (log2-scale) in Col-0 versus *ago1-25*. Black data points and line show the mean oligoribonucleotide standards fit with a linear model. **(C)** Volcano plot of fold change (mean [n = 3] RPM in *ago1-25* divided by mean [n = 3] RPM in wild type, log2-scale) versus absolute SAM score, *d*(*i*). **(D)** Scatter plot of mean (n = 3) RPM in Col-0 (log2-scale) versus absolute SAM score, *d*(*i*). In **B–D**, SAM-seq significant data points are color-coded red (overrepresented) and green (underrepresented).

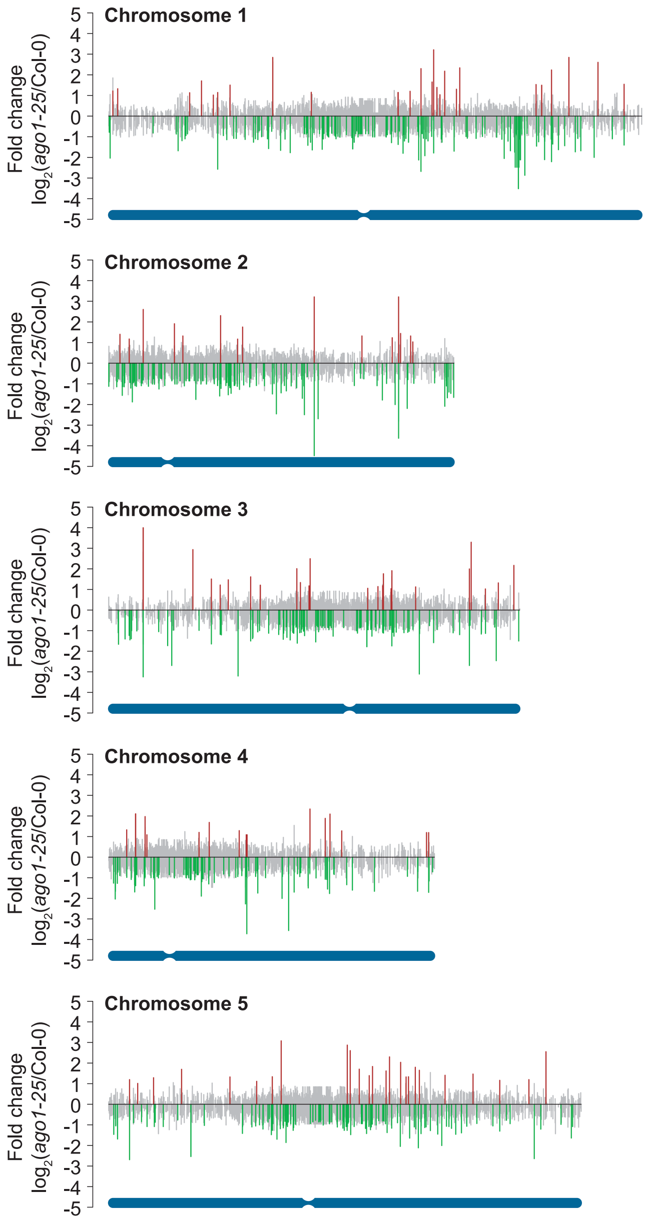
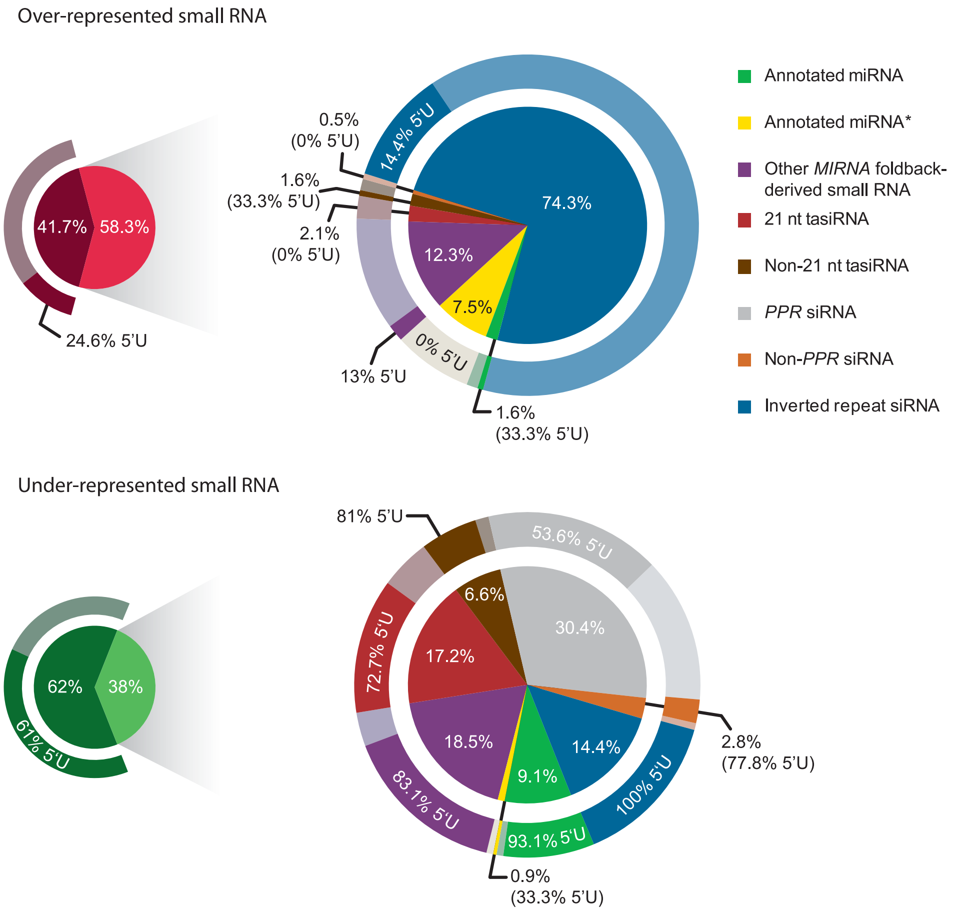


Figure 2.5. Genome-wide view of differentially expressed small RNA identified by SAM-seq.

Fold change (mean [n = 3] RPM in *ago1-25* divided by mean [n = 3] RPM in wild type, log2-scale) plotted for each unique small RNA analyzed by SAM-seq (n = 359,447) by position across each *Arabidopsis* chromosome. Small RNAs are color-coded gray (nonsignificant), red (significantly overrepresented), and green (significantly underrepresented). Chromosomes are illustrated in blue at the bottom of each graph with centromeres marked.



Over-represented small RNA

Under-represented small RNA

Figure 2.6. Profile of over- and underrepresented small RNA in *ago1-25* identified by SAM-seq.

Small pie charts (left) represent the percentage of total over- or underrepresented small RNA in *ago1-25* that were (light red or light green) or were not (dark red or dark green) categorized as previously annotated miRNA, annotated miRNA*, other *MIRNA* foldback-derived small RNA, 21-nucleotide tasiRNA, other *TAS* locus-derived small RNA, RDR6/DCL4-dependent *PPR*-derived siRNA (Howell et al., 2007), RDR6/DCL4-dependent siRNA derived from non-*PPR* genes and inverted repeat-derived siRNA. Breakdowns of categorized small RNA are expanded in the larger pie charts. The larger pie charts (right) show the percentage of small RNA in each category listed in the key (top right). Outer rings around each pie chart show percentage of small RNA from each category that possess a 5'U.

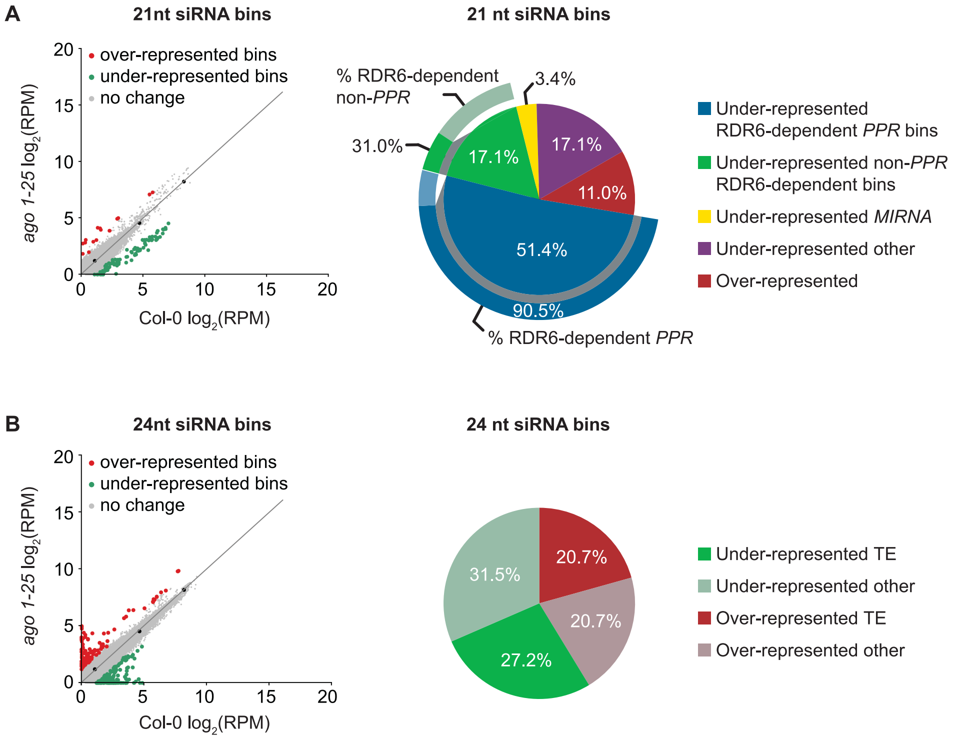


Figure 2.7. SAM-seq analysis of differentially expressed small RNA-containing bins between wild type Col-0 and *ago1-25*.

Analysis of **(A)** 21- and **(B)** 24-nucleotide siRNA bins (1000-nucleotide windows, 200-nucleotide scroll). Scatter plots (left) of the mean (n = 3) RPM for small RNA-containing bins in Col-0 versus *ago1-25*. Red and green data points show statistically significant over- and underrepresented bins, respectively. Black data points and lines show the oligoribonucleotide standards fit with a linear model. Gray data points show nonsignificant bins. Pie charts (right) show the percentage of bins identified in each category. The outer arcs represent the percentage of *PPR* and non-*PPR* genes that were identified as significant 21-nucleotide siRNA bins by SAM-seq, and that were identified in previous analyses of RDR6/DCL4-dependent siRNA clusters (Axtell et al., 2006; Howell et al., 2007).

SUPPLEMENTAL DATA

The following materials are available in the online version of this article (http://rnajournal.cshlp.org).

Supplemental Table 1. Over- and under-represented unique small RNA identfied by SAM-seq (Col-0 vs *ago1-25*).

Supplemental Table 4. Over- and under-represented 21-nucleotide siRNA bins identified by SAM-seq (Col-0 vs *ago1-25*).

Supplemental Table 5. Over- and under-represented 24-nucleotide siRNA bins identified by SAM-seq (Col-0 vs *ago1-25*).

Supplemental Table 6. Delta table for SAM-seq analysis of unique small RNA in wild type Col-0 and *ago1-25*.

Supplemental Table 7. Delta table for SAM-seq analysis of 21-nucleotide siRNA bins in wild type Col-0 and *ago1-25*.

Supplemental Table 8. Delta table for SAM-seq analysis of 24-nucleotide siRNA bins in wild type Col-0 and *ago1-25*.

Supplemental Table 9. Miss-rate table for SAM-seq analysis of 21-nucleotide siRNA bins in wild type Col-0 and *ago1-25*.

Supplemental Table 10. Miss-rate table for SAM-seq analysis of 24-nucleotide siRNA bins in wild type Col-0 and *ago1-25*.

Additional CASHX Information

cashx_formatDB and cashx_formatDBmem generate a database with a 4 and 12 nucleotide index length, respectively. The index size can be changed once a database has been created using the program cashx_translateDB. The index size of a database is checked using the program cashx_checkDB. The input for cashx_formatDB and cashx_formatDBmem must be in FASTA format with no line returns within the sequence. To accommodate this, we provide the Perl script fasta_stitcher.pl that removes line returns in the sequence lines of a FASTA-formatted file.

CASHX databases are searched using the program cashx_searchDB. A FASTA-formatted sequence file is provided to cashx_searchDB along with the name of a CASHX database. We provide two scripts for generating input files. The first Perl script, eland_2_fasta.pl, converts files in ELAND format to FASTA format. The second script, eland_smallRNA_parser.pl, will also convert a file from ELAND format to FASTA format, but the script will also parse adaptor sequences off of small RNA sequences or other sequences that contain amplicon adaptor sequence.

In addition to the CASHX mapping tools, the CASHX package contains the entire pipeline used to parse sequences from adaptor sequences (both 5' barcoded and 3' adaptors), quantify numbers of reads per unique parsed sequence, map sequences to a reference genome and store run data in a MySQL database. The CASHX suite and full documentation is available for download at http://jcclab.science.oregonstate.edu/?q=node/view/54596.

Identification of novel miRNA

Since previous analyses demonstrated that AGO1 has preference for miRNA and tasiRNA with a 5'U, we reasoned that some novel miRNA might be represented in the set of small RNA significantly affected by the *ago1-25* mutant (Fahlgren et al., 2009). We searched the list of significant small RNA for sequences derived from the arms of potential foldbacks that produced small RNA primarily from one strand (strand bias was at least 5 to 1 for the miRNA strand in wild type samples). Predicted transcripts from genomic regions were subjected to RNAfold (Hofacker, 2003) with free energy calculations simulated at 22°C, with no GU base pairs allowed to close helices and the "dangling end" energy check disabled (Figure S2.2). We found three candidate miRNA from four loci (Table S2.2). In the case of *MIR1886* and *MIR2111a/b*, the significantly affected small RNA was also the most abundant species derived from the foldback, and was therefore considered to be the miRNA (Table S2.2). For *MIR2112*, the significantly affected sequence was not the most abundant small RNA species. Instead, a pair of sequences that form a canonical miRNA/miRNA* duplex were nearly equally abundant and were given the -5p and -3p designations (Table S2.2). Targets were predicted for each miRNA as in Fahlgren et al., (2007) (Table S2.2).

Additionally, we found that miR1886.2, which was also identified as a candidate by Rajagopalan et al., (2006), was derived from the recently identified *MIR1886* foldback, but was offset by nine nts relative to annotated miR1886.1 (German et al., 2008; Figure S2.2). miR1886.2 was the most abundant small RNA derived from *MIR1886*, although German et al., (2008) showed that miR1886.1 likely functions to guide cleavage of target mRNA. Given the abundance of both miR1886.1 and miR1886.2, the presence of miRNA* for each in the available databases (ASRP, http://asrp.cgrb.oregonstate.edu/db and Arabidopsis SBS, http://mpss.udel.edu/at_sbs) and the fact that they are offset by nine nts (Figure S2.2), we propose that these small RNA represent offset variants like those seen from *MIR161* (Allen et al., 2004).

Table S2.1. Miss rates for SAM-seq analysis of unique small RNA from wild type Col-0 and *ago1-25*

Quantiles[a]	**Cutpoints [*d(i)*]**[b]	**Miss Rate (%)**[c]
0 -> 0.05	-0.75 -> -0.328	39.11
0.05 -> 0.1	-0.328 -> -0.244	16.04
0.1 -> 0.15	-0.244 -> -0.197	17
0.15 -> 0.2	-0.197 -> -0.159	7.77
0.2 -> 0.25	-0.159 -> -0.13	7.86
0.25 -> 0.75	-0.13 -> 0.098	0.86
0.75 -> 0.8	0.098 -> 0.125	0
0.8 -> 0.85	0.125 -> 0.159	0
0.85 -> 0.9	0.159 -> 0.207	0
0.9 -> 0.95	0.207 -> 0.281	0
0.95 -> 1	0.281 -> 0.835	0

[a] Quantiles of small RNA called not significant by SAM-seq.

[b] Ranges of the relative difference score, *d(i)*, for small RNA in the given quantile range.

[c] Estimated false negative rate for the given range of *d(i)*.

Table S2.2. Novel *A. thaliana* miRNA

Name	miRNA sequence	Foldback position[a]	Strand	Average RPM[b]		Predicted target family[d]	Predicted targets[d]
				Col-0	*ago1-25*		
miR1886.2	UGAGAUGAAAUCUUUGAUUGG	2:15618940-15619066	1	19.3	4.2	Unknown, bZIP transcription factor, fasciclin-like arabinogalactan	At2g38770 [3.5], At3g10800 [3.5], At4g08940 [3.5], At5g06390 [3.5]
miR1886.2*	AAUUAAAGAUUUCAUCUUACU			0.1	0		
miR2111a	UAAUCUGCAUCCUGAGGUUUA	3:2854290-2854448	1	3.2	0.5	F-box, calcineurin-like phosphoesterase, 1-aminocyclopropane-1-carboxylate oxidase	At3g27150 [2], At1g07010 [3], At2g19590 [3.5]
miR2111a*	GTCCTCGGGATGCGGATTACC			0[c]	0		
miR2111b	UAAUCUGCAUCCUGAGGUUUA	5:400355-400521	1	3.2	0.5	F-box, calcineurin-like phosphoesterase, 1-aminocyclopropane-1-carboxylate oxidase	At3g27150 [2], At1g07010 [3], At2g19590 [3.5]
miR2111b*	ATCCTCGGGATACAGTTTACC			0[c]	0		
miR2112-5p	CGCAAAUGCGGAUAUCAAUGU	1:234006-234159	-1	2.4	13.7	Pseudo-repsonse regulator (APRR8), RS zinc knuckle	At4g00760 [1], At2g37340 [3.5]
miR2112-3p	CUUUAUAUCCGCAUUUGCGCA			2.6	0.5	Pseudo-repsonse regulator (APRR8)	At4g00760 [2]

[a] Chromosome:start-end.

[b] Average reads per million (RPM) for three replicates of wild type Col-0 and three replicates of *ago1-25*.

[c] miRNA* sequenced in another library (Rajagopalan et al., 2006).

[d] Computationally predicted miRNA targets with target scores of 3.5 or less. Targets were predicted as in Fahlgren et al. (2007).

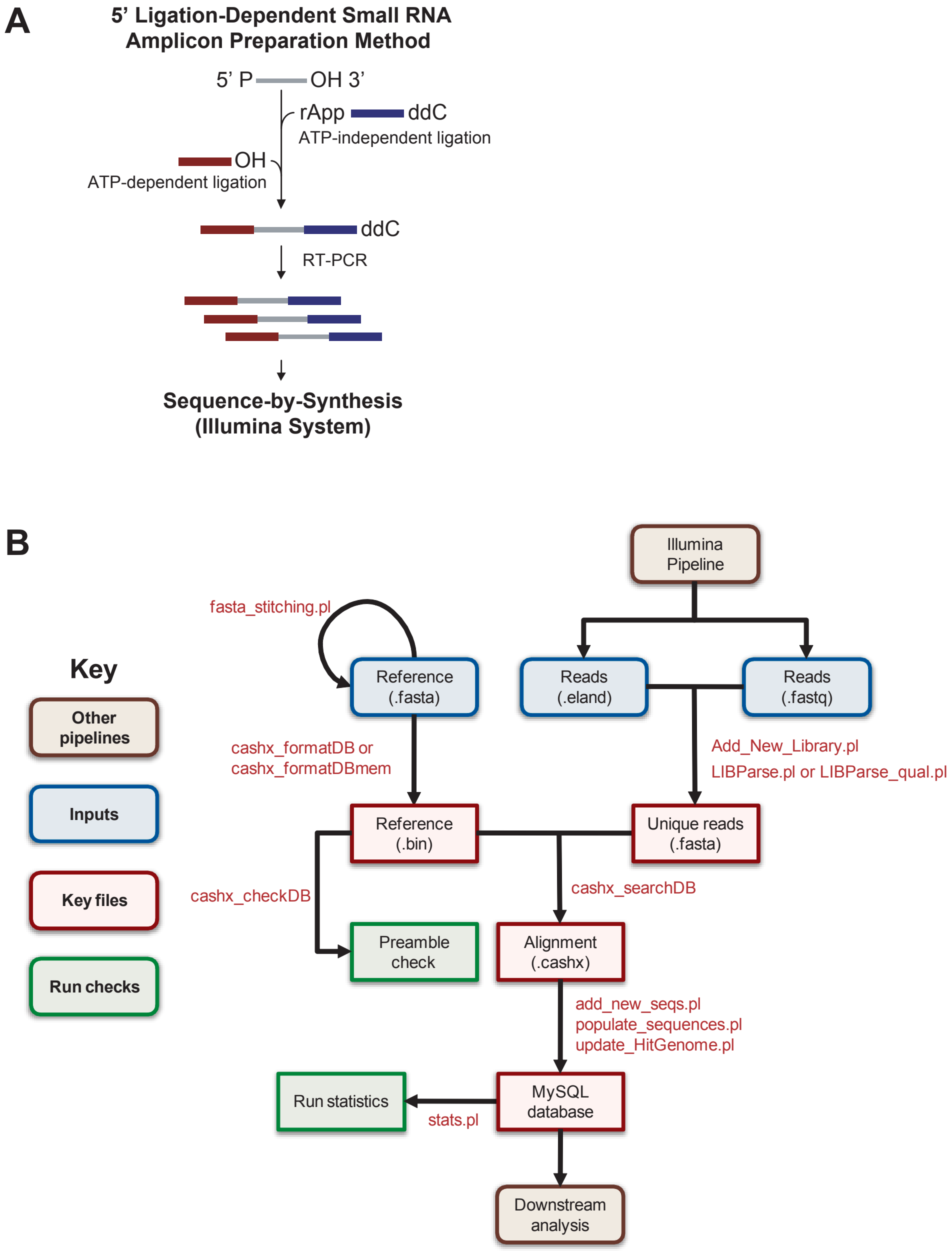


Figure S2.1. Methods flowcharts.
(A) Flowchart for preparation of 5’ ligation-dependent amplicons for SBS. The full protocol for preparing 5’ ligation-dependent small RNA amplicons is available for download at http://jcclab.science.oregonstate.edu/?q=node/view/54596. **(B)** Workflow diagram for processing and mapping of SBS reads using CASHX. Program names for each step are listed in red.

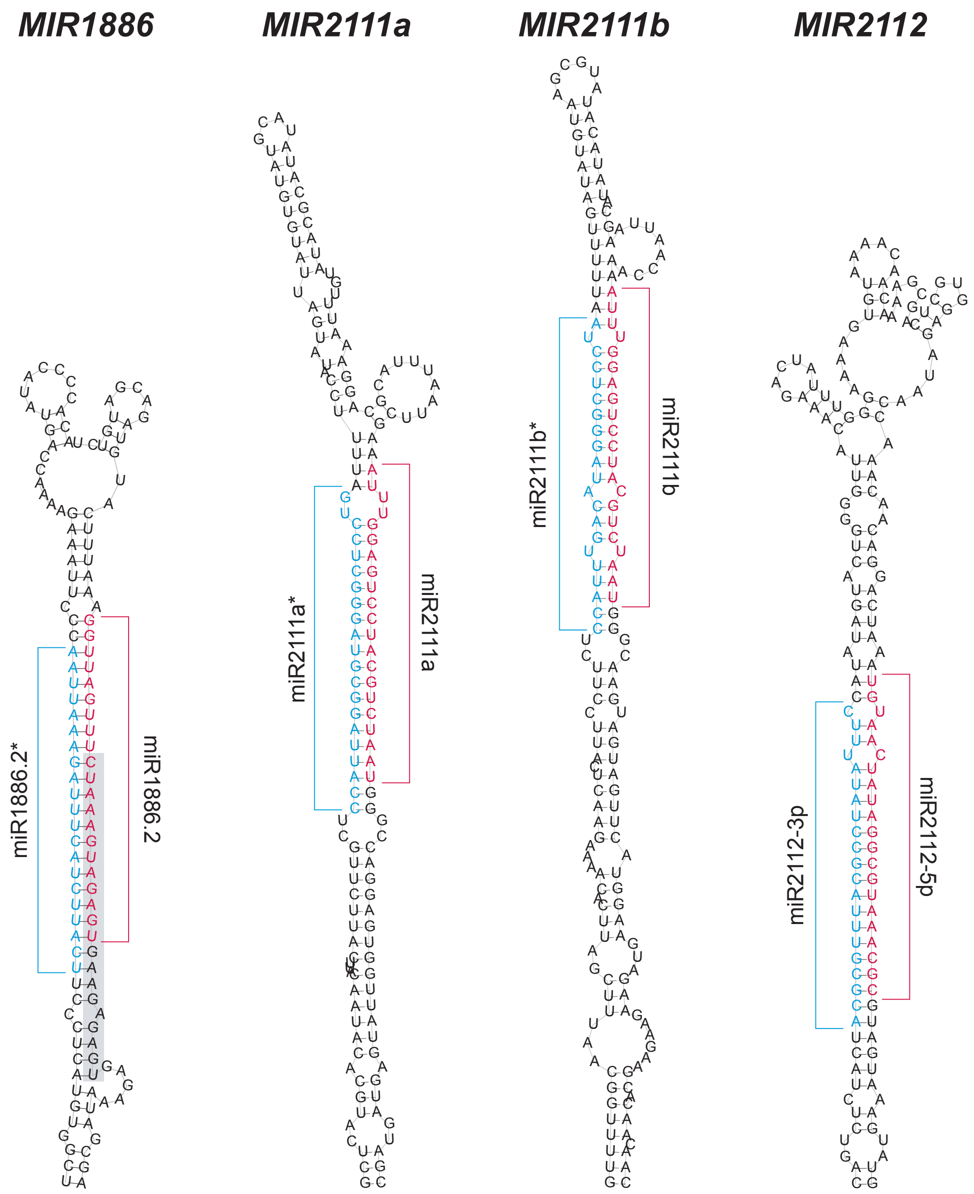


Figure S2.2. New *MIRNA* foldbacks.
Foldback structures were generated using RNAfold. For *MIR1886*, *MIR2111a* and *MIR2111b*, miRNA are colored red and miRNA* are colored in blue. For *MIR2112*, the -5p and -3p nomenclature is used because small RNA from both arms of the foldback were sequenced approximately the same number of times. In the *MIR1886* foldback, previously annotated miR1886.1 is shaded (grey box).

High-throughput Sequencing of *Arabidopsis* microRNAs: Evidence for Frequent Birth and Death of *MIRNA* Genes

Noah Fahlgren, Miya D. Howell, Kristin D. Kasschau, Elisabeth J. Chapman, Christopher M. Sullivan, Jason S. Cumbie, Scott A. Givan, Theresa F. Law, Sarah R. Grant, Jeffery L. Dangl, and James C. Carrington

PLoS ONE, Public Library of Science
1160 Battery Street,
Koshland Building East, Suite 100
San Francisco, California 94111
Volume 2, pages e219

SUMMARY

In plants, microRNAs (miRNAs) comprise one of two classes of small RNAs that function primarily as negative regulators at the posttranscriptional level. Several *MIRNA* genes in the plant kingdom are ancient, with conservation extending between angiosperms and the mosses, whereas many others are more recently evolved. Here, we use deep sequencing and computational methods to identify, profile and analyze non-conserved *MIRNA* genes in *Arabidopsis thaliana*. 48 non-conserved *MIRNA* families, nearly all of which were represented by single genes, were identified. Sequence similarity analyses of miRNA precursor foldback arms revealed evidence for recent evolutionary origin of 16 *MIRNA* loci through inverted duplication events from protein-coding gene sequences. Interestingly, these recently evolved *MIRNA* genes have taken distinct paths. Whereas some non-conserved miRNAs interact with and regulate target transcripts from gene families that donated parental sequences, others have drifted to the point of non-interaction with parental gene family transcripts. Some young *MIRNA* loci clearly originated from one gene family but form miRNAs that target transcripts in another family. We suggest that *MIRNA* genes are undergoing relatively frequent birth and death, with only a subset being stabilized by integration into regulatory networks.

INTRODUCTION

Eukaryotes possess RNA silencing systems to regulate or suppress a range of genes, genetic elements, and viruses (Meister and Tuschl, 2004). Regulation by RNA silencing can occur at either the transcriptional or posttranscriptional level, although in both cases, silencing is associated with formation of small RNA classes with typical sizes of 21 and 24 nucleotides (nts) (Baulcombe, 2004; Meister and Tuschl, 2004; Tomari and Zamore, 2005). Small RNA biogenesis occurs from perfect or near-perfect double-stranded RNA (dsRNA) that arises by synthesis of self-complementary foldbacks, by bidirectional transcription, or through the activity of RNA-dependent RNA polymerases (RDR) (Baulcombe, 2004; Meister and Tuschl, 2004; Tomari and Zamore, 2005). Processing of self-complementary foldback or dsRNA precursors to small RNA duplexes is catalyzed by complexes containing DICER [or DICER-LIKE (DCL)] proteins and dsRNA-binding proteins (Baulcombe, 2004; Meister and Tuschl, 2004; Tomari and Zamore, 2005). Single-stranded small RNAs then associate with ARGONAUTE (AGO) proteins in effector complexes (Hall, 2005; Hammond, 2005). For transcriptional silencing, effector complexes associate (directly or indirectly) with factors controlling repressive chromatin, including DNA methylation and histone modification enzymes (Noma et al., 2004; Chan et al., 2005; Tran et al., 2005). Posttranscriptional silencing effector complexes can mediate irreversible cleavage, translational repression, or subcellular redirection of target transcripts (Baulcombe, 2004; Meister and Tuschl, 2004; Liu et al., 2005; Tomari and Zamore, 2005; Chu and Rana, 2006).

The expanse of genetic information regulated posttranscriptionally by small RNAs is potentially large in animals and plants (Jones-Rhoades et al., 2006; Mallory and Vaucheret, 2006; Rajewsky, 2006). In humans, for example, computational and indirect experimental evidence indicates that miRNAs regulate expression of up to 1/3 of all genes (Lewis et al., 2003; John et al., 2004; Farh et al., 2005; John et al., 2006). In plants, far fewer mRNAs are directly regulated by miRNAs, although the direct and indirect consequences of miRNA-directed regulation are significant (Jones-Rhoades et al., 2006; Mallory and Vaucheret, 2006). This is due to the roles of a large proportion of plant miRNA target transcripts that encode transcription factors required for normal growth, development, hormone response, meristem functions and stress responses (Rhoades et al., 2002; Jones-Rhoades and Bartel, 2004; Jones-Rhoades et al., 2006; Mallory and Vaucheret, 2006). Approximately 21 families of *Arabidopsis* miRNAs and their respective targets are conserved in Rice and/or Poplar (Jones-Rhoades et al., 2006). Additionally, there is a growing recognition of significant numbers of miRNAs not conserved in Rice or Poplar, many of which likely arose in the recent evolutionary past (Allen et al., 2004; Lindow and Krogh, 2005; Lu et al., 2005b; Lu et al., 2006; Maher et al., 2006). In some cases, these *MIRNA* loci formed through inverted duplication events that yielded transcripts with self-complementary foldback structure (Allen et al., 2004). In fact, the genomes of *Arabidopsis* and other plants contain a wide diversity of non-conserved, local inverted duplications that yield small RNA populations ranging from highly uniform, DCL1-dependent miRNAs (e.g. *MIR163*) to heterogeneous collections of bi-directional short interfering RNAs (siRNAs) formed by multiple DCLs (Allen et al., 2004; Slotkin et al., 2005; Henderson et al., 2006; Lu et al., 2006).

Deep sequencing methods now provide a rapid way to identify and profile small RNA populations in different plants, mutants, tissues, and at different stages of development. We, and others, have used high-throughput pyrosequencing to analyze small RNAs across the *Arabidopsis* genome in wild type and silencing-defective mutants (Lu et al., 2005a; Axtell et al., 2006; Henderson et al., 2006; Lu et al., 2006; Kasschau et al., 2007). In this paper, we identify and analyze non-conserved and recently evolved *Arabidopsis* miRNAs. The data reveal a relatively large number of miRNAs that are so far unique to *Arabidopsis*, and suggest that many miRNAs are spawned and lost frequently during evolution.

RESULTS AND DISCUSSION

New miRNAs and miRNA target transcripts

Small RNA populations from wild type (Col-0) plants, from *dcl* (*dcl1-7*, *dcl2-1*, *dcl3-1* and *dcl4-2*) and from *rdr* (*rdr1-1*, *rdr2-1* and *rdr6-15*) mutant plants, as well as from several tissue types of wild type and *rdr6-15* mutant plants, were sequenced using picoliter-scale pyrosequencing (Margulies et al., 2005; Kasschau et al., 2007). This yielded quantitative profiling data for several classes of miRNAs and siRNAs (Kasschau et al., 2007). Procedures for sequencing in a multiplexed format, normalization across samples, normalization for multi-locus small RNAs, and viewing at the *Arabidopsis* Small RNA Project database (ASRP) (http://asrp.cgrb.oregonstate.edu/db/) were described (Kasschau et al., 2007). In addition to the populations analyzed previously, small RNAs were profiled from non-infected leaves (15,826 reads), or leaves that were infected for 1 hr (18,368 reads) or 3 hr (10,363 reads) by *Pseudomonas syringae* pv. tomato (DC3000*hrcC*). In total, 218,575 unique small RNAs (663,312 loci), represented by 470,426 reads, were used in this study.

A computational analysis to identify new *MIRNA* genes was done using a protocol similar to that of Xie et al., (2005b) (Figure 3.1A). All small RNAs from the ASRP database (Set1) were used (Figure 3.1A). Briefly, all loci from Set1 that yielded at least two reads (Set2) were subjected to Repeatmasker (Jurka et al., 2005) and bidirectional small RNA cluster filters to eliminate siRNAs from repeat sequence classes. Small RNAs that differed at their termini by up to four nucleotide positions were consolidated and passed through a self-complementary foldback screen with settings as described (Xie et al., 2005b). Small RNAs from coding sequences and complex small RNA clusters, or that were not 20–22 nucleotides in length, were eliminated, yielding 228 loci (Set3). These included 91 of 102 (false negative rate = 0.11) rigorously characterized *MIRNA* genes used as a rule-development and reference set (Table 3.1) (Xie et al., 2005b). Failure to identify miR398a, miR399a, miR399d, miR399e, miR399f and miR447c was due to a lack of sequence reads. Known miRNAs also failed due to length (miR163), an incorrectly predicted foldback (miR164b and miR167d) or because the *MIRNA* gene yielded bidirectional small RNAs (miR156d and miR161). Small RNAs from known *MIRNA* genes were then removed from Set3, leaving 79 unique small RNA loci (Figure 3.1A). Two loci that had each failed at one filter step were reclaimed manually due to their abundance and their dependence on DCL1 (http://asrp.cgrb.oregonstate.edu/db/). The 81 loci (Set4) were then subjected to a detailed foldback analysis where small RNAs from the same foldback were consolidated into a single prospective *MIRNA* locus. A particularly useful, although not absolute, feature was detection of miRNA* sequences, which arise from the opposite foldback arm during DCL1-mediated processing. Detection of miRNA* sequences reveals the functionality of the predicted foldback. miRNA* sequences were detected at relatively low abundance for most reference *MIRNA* families (Table 3.1).

Thirty-nine loci (Set5) beyond the reference set emerged as prospective miRNAs, including 25 for which miRNA* sequences were detected (Table 3.2). Eight miRNAs from Set5 were detected recently as miRNAs by Lu et al., (2006). Only one sequence, corresponding to *Populus trichocarpa* miR472 (Lu et al., 2005b), was conserved in Poplar based on sequence/foldback similarity and target conservation. None of the new sequences were conserved in Rice. In several cases (miR830, miR833, miR851, miR861, miR862, miR863, miR864, miR865 and miR866), due to the ratio of small RNA sequences from both the 5p and 3p strands of the foldback, it was not clear which small RNA should be designated as the miRNA. In such cases, both the 5p and 3p strand sequences are designated (Table 3.2). In total, at least 70 *MIRNA* families were detected in at least one sequenced population (Tables 3.1 and 3.2, and references therein). Most miRNAs were lost or underrepresented in the *dcl1-7* mutant (http://asrp.cgrb.oregonstate.edu/db/).

Target RNAs were predicted from both transcript and EST databases for the 39 miRNAs in Set5 by the method of Allen et al., (2005), which is based on additive, position-dependent mispair penalties. Seventy-eight or 69 of 85 validated, reference miRNA targets were predicted with a threshold score of 4 or 3.5, respectively (false negative rate = 0.08 or 0.19; Figure 3.1A). A total of 142 targets were predicted for the Set5 miRNAs, using the more conservative 3.5 score threshold (Table 3.2). These included targets predicted for both 5p and 3p strands for eight *MIRNA* loci. The 3.5 score threshold was used to maintain specificity, although at the cost of sensitivity, in the target prediction algorithm.

Top-scoring predicted targets for most miRNAs from Set5, and two that were not found in this study (miR778 and miR781) (Lu et al., 2006), were tested using a standard 5’RACE analysis to detect cleavage events at predicted sites opposite nucleotide 10 from the 5’ end of the small RNA. The 5’RACE assays were done using two gene-specific primer sets with RNA from whole seedling and inflorescence tissues. Thirteen targets for 13 miRNAs were validated, although evidence supporting miR775- and miR859-guided cleavage (At1g53290 and At3g49510, respectively) was weak (Figure 3.1B, Table 3.2). Evidence for miR846- and miR844-guided cleavage should also be interpreted cautiously, as targets were not cleaved at the canonical position. These were considered validated because both miRNAs had multiple sequenced variants that had predicted target sites encompassing the observed cleavage sites (Figure 3.1B; http://asrp.cgrb.oregonstate.edu/db/). Additionally, targets for miR472 and miR774 were validated recently (Lu et al., 2006). Each target-validated miRNA was represented by at least ten reads or had a predicted miRNA* sequence represented with at least one read in the database, except for miR774 and miR778 which were not sequenced here (Table 3.2). Twenty-two predicted miRNAs that were sequenced at least two times, or that had miRNA* sequences represented in the ASRP or MPSS databases (Nakano et al., 2006), failed at the target validation step (Table 3.2).

The relatively high proportion of target validation failures could be due to erroneous target predictions, low-abundance targets or miRNA-guided cleavage products, or low-abundance miRNAs with limited or no activity. Additionally, for 13 miRNAs, no targets were predicted at a score threshold of 3.5 (Table 3.2). While it is possible that many of these function to guide cleavage of unpredicted target RNAs, it is also possible that miRNAs exist without actual targets (see below).

Several new target families are worth noting. First, at least two targets are clearly under negative regulation by miRNAs in inflorescence tissue. Transcripts from *AGL16* (At3g57230, MADS- box) and *MYB12* (At2g47460, R2R3-MYB family) genes are targeted by miR824 and miR858, respectively, and are each up- regulated by 2–4 fold in *dcl1-7* and *hen1-1* mutants (Table 3.2, Figure 3.2A). Functions for these specific transcription factors are not known. Second, the transcript for the cation/hydrogen antiporter gene *CHX18* is targeted by both miR780 and miR856 (Table 3.2, Figure 3.1B). Dual targeting of the *CHX18* (At5g41610) transcript may lead to secondary, 21-nucleotide siRNAs that arise through the RDR6/DCL4-dependent pathway ((Peragine et al., 2004; Vazquez et al., 2004; Allen et al., 2005; Gasciolli et al., 2005; Xie et al., 2005a; Yoshikawa et al., 2005; Axtell et al., 2006); data not shown). miR856 may also target an unrelated transcript encoding the efflux protein ZINC TRANSPORTER OF ARABIDOPSIS THALIANA1 (ZAT1). Third, miR857 was validated to target the *LAC7* (At3g09220) transcript, which encodes a laccase family protein with predicted multicopper oxidase function. This represents the third miRNA family to target the laccase gene family (Jones-Rhoades and Bartel, 2004; Sunkar and Zhu, 2004; Schwab et al., 2005). In fact, at least seven gene or domain families (encoding MYB, PENTATRICOPEPTIDE REPEAT (PPR), AUXIN RESPONSE FACTOR (ARF), AGO, F-box domain, kinase and LAC proteins) are now known to be targeted by multiple miRNA families (Jones-Rhoades et al., 2006).

Conserved vs. non-conserved miRNAs

Twenty-two and 20 *MIRNA* families (Tables 3.1 and 3.2) are conserved in Poplar and Rice, respectively. Conserved families in *Arabidopsis* were designated by having identical or related (three or fewer nucleotide substitutions) sequences, and at least one conserved target transcript, in either Poplar or Rice. A series of general and functional comparisons between conserved and non-conserved families was done. Whereas 19 of 22 conserved *Arabidopsis* miRNA families were represented at multiple loci, only three (*MIR158*, *MIR447* and *MIR845*) of 48 non-conserved miRNAs were members of multigene families (Figure 3.2B). Expansion of multigene *MIRNA* families has occurred through tandem and segmental duplications, as well as polyploidization events (Allen et al., 2004; Guddeti et al., 2005; Jiang et al., 2006). The preponderance of single gene families among the non-conserved miRNAs is consistent with recent evolutionary derivation.

Functions of conserved miRNA target genes, as a group, are less diverse than functions for non-conserved miRNA target genes (only validated or high-confidence targets were compared). This is due primarily to the high proportion of target mRNAs encoding transcription factors for conserved families (Table 3.1 and 3.2, Figure 3.2C). As pointed out clearly before (Rhoades et al., 2002; Jones-Rhoades and Bartel, 2004), the vast majority of transcription factor families targeted by conserved miRNAs participate in developmental pathways, including those specifying meristem functions, organ polarity, cell division control, organ separation and hormone responses. Non-conserved miRNA target genes encode a broad range of proteins, including a limited number of transcription factors (Table 3.2, Figure 3.2C). Interestingly, several non-conserved miRNAs (miR161 and miR400) target transcripts from a clade within the large *PPR* family (Rhoades et al., 2002; Sunkar and Zhu, 2004). Another non-conserved miRNA, miR173, targets tasiRNA primary transcripts (*TAS1* and *TAS2*) (Allen et al., 2005), which in turn yield siRNAs that also target several of the miR161- and miR400-targeted *PPR* transcripts ((Axtell et al., 2006); data not shown).

To what extent do the non-conserved miRNAs negatively regulate target genes? The sensitivity of target genes of conserved and non-conserved miRNAs was analyzed by transcript profiling in wild type (Col-0 and La-*er*) plants and mutant plants with general miRNA deficiencies (*dcl1-7* and *hen1-1*). To avoid biasing the analysis with miRNAs that target disproportionately high numbers of target mRNAs, such as miR161 and *PPR* target gene family members, the number of genes analyzed for each miRNA family was limited to two validated targets, or two predicted targets with the lowest scores. For miRNAs that have been shown to target multiple gene families, such as miR395, two targets from both gene families were analyzed. A scatterplot of fold-change in *dcl1-7* and *hen1-1* mutants for each target was generated. As shown previously (Allen et al., 2005), conserved miRNA target transcripts displayed a generally elevated pattern in both mutants (Figure 3.2A). In contrast, most target transcripts of non-conserved miRNAs were clustered around the origin, indicating that most were insensitive to either the *dcl1-7* or *hen1-1* mutation (Figure 3.2A). The exceptions that were affected at levels of 1.6-fold or greater in either mutant included transcripts from *AGL16*, *MYB12* and a *PPR* gene (At1g63130), which were targeted by miR824, miR858 and miR161.1, respectively (Figure 3.2A). While some of these data may be skewed by tissue-specific expression patterns of miRNAs and target genes, the general trend for non-conserved miRNAs having fewer effects on target transcript levels is clear. We suggest that, in contrast to the vast majority of conserved miRNAs, a high proportion of the non-conserved miRNAs are not integrated as dominant factors within regulatory networks.

Recent evolution of non-conserved *MIRNA* loci

Previously, we identified a number of small RNA-generating loci with the potential to yield transcripts with self-complementary foldback potential and with extensive similarity to protein-coding gene family sequences (Allen et al., 2004). Whereas the majority of these loci yield complex populations of siRNAs, two (*MIR161* and *MIR163*) yield functional, non-conserved miRNAs (Allen et al., 2004). This led to the idea that aberrant replication/recombination or transposition events from expressed gene sequences can spawn new small RNA-generating loci with the potential to evolve into *MIRNA* genes that regulate members of the originating family.

This idea was tested more rigorously with the expanded set of *MIRNA* loci. Foldback sequences for each *MIRNA* locus (Tables 3.1 and 3.2) were used in FASTA searches against *Arabidopsis* transcript and gene databases (Figure 3.3A). Nearly all foldback sequences from conserved miRNAs had hits with non-significant E-values greater than 0.05 (Figure 3.3B). In contrast, 19 of 48 non-conserved miRNA foldback sequences had at least one hit with an E-value lower than 0.05 (Figure 3.3B). Similarity or complementarity was detected on both 5' and 3' arms containing miRNA or miRNA* sequences.

It is important to recognize that *MIRNA* loci contain two regions – the miRNA and miRNA-complementary region (largely overlapping with the miRNA*) – with relatively high levels of complementarity or similarity to target genes. To eliminate the potential misleading influence of these sequences, which may be under selection due to the requirement for complementarity between miRNAs and their targets, on the similarity test, each foldback arm with hits (E<0.05) in the FASTA search was analyzed independently with and without the miRNA or miRNA-complementary sequences. The top four FASTA hits (Figure 3.3A) were aligned with the intact or deleted arms using a global sequence alignment method (NEEDLE). *MIR163* and *MIR161* were analyzed in detail previously (Allen et al., 2004) and were included here as controls. Two foldback arm sets from *MIR161*, with miR161.1/miR161.1-complementary and the overlapping miR161.2/miR161.2-complementary sequences deleted independently, were analyzed. Sixteen foldbacks, including the *MIR161* and *MIR163* controls, contained at least one arm with similarity or complementarity (NEEDLE score with p<0.001) to one or more genes when an alignment was done with both intact and deleted arms (Figure 3.3C, Table S3.3). For all hits with the deleted arms, the top-score gene alignments identified sequences that were immediately flanking the region with similarity or complementarity to miRNA regions (Figure 3.4). Further, in all cases in which multiple genes were hit with intact and deleted arms, the genes were closely related members of a single family. Thus, over 30% of the non-conserved *MIRNA* loci show evidence of common origin with specific genes or gene families.

Interestingly, the two *MIR824* arms aligned best with distinct regions within one gene (*AGL16*). The miR824 arm was complementary to a region from exon VIII, and the miR824* arm was most similar to a duplicated region located within intron III. The two arms from *MIR846* were most similar to two adjacent but distinct regions of jacalin domain-containing genes or pseudogenes (At1g57570 and At1g61230). It is likely, therefore, that some recently evolved *MIRNA* loci arose through juxtapositioning of sequences from two related duplicated sequences.

The protein-coding genes with extended *MIRNA* foldback arm similarity were analyzed in more detail. The intact foldback arms from 13 of the 16 gene-similar *MIRNA* loci were aligned with up to three gene sequences, and alignment quality was displayed using heat maps (Figure 3.4; *MIR161*, *MIR163* and *MIR447c* were not included). This revealed several sequence conservation patterns. For several *MIRNA* arm-gene alignments, including those containing miR778, miR780, miR824 and miR856 sequences, the aligned region containing miRNA sequences was clearly more conserved than the remaining arm segments (Figure 3.4). This suggests that selection may be operating on several of the recently evolved miRNA sequences, and is indirect evidence for functional significance of the miRNA. Indeed, miR824 targets the *AGL16* transcript, which is under miRNA pathway-dependent repression (Figure 3.2A).

For several other *MIRNA* arm-gene alignments, such as those containing miR862, miR447a and miR853, the miRNA-region of the alignment is similar to, or weaker than, the alignment containing flanking arm sequences (Figure 3.4). In fact, among the 16 *MIRNA*-similar gene sets identified in this analysis (Table S3.3), only seven represented sets that corresponded to the best predicted or validated miRNA-target pair (Figures 3.3 and 3.5). In other words, nine of the miRNAs from loci with similarity to protein-coding genes were predicted to target transcripts from different genes (based on best target prediction scores). The target scores for eight of these sets were clearly weak or functionally implausible (Figure 3.5, Table S3.3). Interestingly, miR447a and miR856 were validated to target transcripts in gene families distinct from those similar to the foldback sequences (Figure 3.1B, Table 3.2; (Allen et al., 2005)). The miR447 family appears to have acquired novel target specificity and lost sequence-of-origin specificity, while miR856 may have acquired dual-targeting specificity (Figure 3.5, Table 3.2, Table S3.3).

The picture that emerges from analysis of the non-conserved *Arabidopsis* miRNAs appears to show that new *MIRNA* loci are forming frequently through duplication events. These newly evolved *MIRNA* loci may first pass through a stage in which heterogeneous populations of siRNA-like sequences are generated, especially if the duplicated locus and resulting foldback sequence is large (Allen et al., 2004). This, of course, assumes that the newly spawned sequence is proximal to a functional promoter. Computational evidence indicates that formation of *MIR163* through inverted duplication of SAMT methyltransferase-like sequences also involved duplication of the gene-of-origin promoter (Wang et al., 2006). Given that DCL1 has limited or insignificant activity on perfectly-paired dsRNA, acquisition of DCL1-dependence and subsequent formation of discrete small RNA products from the foldback precursor may require accumulation of drift mutations that result in foldback mispairs. From this point, we envision three evolutionary fates. The first and perhaps most common is continued sequence drift through mutation, decay of both targeting capacity and promoter (if not required for other functions) sequences, and eventual death of the locus. New *MIRNA* genes with neutral effects are predicted to take this route. The second fate is stabilization of the miRNA-generating sequence with specificity for gene-of-origin, or family-of-origin, sequences. This would occur if an advantage is realized when the target gene or genes are brought under negative regulation by the miRNA, and would lead to selection of the miRNA sequence. The miR824-*AGL16* regulatory pair may exemplify this evolutionary track. The third fate is chance acquisition of targeting specificity for a novel target gene or family, followed by stabilization through selection in the event of an advantage. miR856, which may target transcripts from both gene-of-origin (*ZAT1*) and a novel gene (*CHX18*), could conceivably fall into this category, although it should be noted that there are no direct data supporting a functional role for miR856. We postulate that many of the conserved *MIRNA* genes arose through the latter two routes, and have lost sequence relatedness to their genetic origin loci due to drift of foldback arm sequences outside of the miRNA/miRNA-complementary regions (Figure 3.3).

While this paper was under review, an article from Rajagopalan et al., (2006) describing results from deep sequencing of *Arabidopsis* small RNAs (887,000 reads) was published. They identified and named 32 new miRNAs (miR822-miR853), and another 39 candidate miRNAs that were not named. Sixteen of the named miRNAs, and eight of the candidates, were identified as miRNAs in this study (Table 3.2). In this paper, 16 miRNAs (miR845b, miR856-miR870) were named, eight of which were identified only here (Table 3.2). Differences in tissue sampling and sequencing depth likely account for most of the differences in miRNAs identified between the two studies.

MATERIALS AND METHODS

Arabidopsis mutants and microarray samples

Mutant lines for *dcl1-7*, *dcl2-1*, *dcl3-1*, *dcl4-2*, *rdr1-1*, *rdr2-1*, *rdr6-15* and *hen1-1* were described previously (Reinhart et al., 2002; Peragine et al., 2004; Vazquez et al., 2004; Xie et al., 2004; Allen et al., 2005; Xie et al., 2005a). All mutants were in the Col-0 background except for *hen1-1* that is in the La-*er* background. All microarray data were generated using Affymetrix ATH1 arrays and are available at Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2005; http://www.ncbi.nlm.nih.gov/geo/). Col-0 (control for *dcl1-7*) and *dcl1-7* data were from experiments described in Xie et al., (2005a) (GEO accession GSE3011, samples GSM65938, GSM65939, GSM65940, GSM65941, GSM65942 and GSM65943), and La-*er* (control for *hen1-1*) and *hen1-1* data were from Allen et al., (2005) (GEO accession GSE2473, samples GSM47014, GSM47015, GSM47016, GSM47034, GSM47035 and GSM47036).

Small RNA libraries and ASRP database

Small RNA libraries and database construction were described by Kasschau et al., (2007). Briefly, small RNA analysis for wild type (Col-0), *dcl1-7*, *dcl2-1*, *dcl3-1*, *dcl4-2*, *rdr1-1*, *rdr2-1*, and *rdr6-15* inflorescence tissue was done by picoliter-scale pyrosequencing [454 Life Sciences (Margulies et al., 2005)]. Small RNA preparations from Col-0 and *rdr6-15* whole seedling, and leaf samples of Col-0 that were either uninoculated or inoculated by *P. syringae* pv. tomato (DC3000*hrcC*) for 1 hr and 3 hr, were also analyzed. Methods for normalization of reads were described previously (Kasschau et al., 2007). All small RNA sequences are available for download GEO accession GSE6682 and the ASRP Database (http://asrp.cgrb.oregonstate.edu/db/).

miRNA identification and target prediction

A set of computational filters based on those developed by Xie et al., (2005b) were used to identify new miRNAs from among sequences in the ASRP database (http://asrp.cgrb.oregonstate.edu/db/), but with the following modifications. First, only small RNAs that were represented by two or more reads were considered. Second, small RNAs arising from repeat elements (Jurka et al., 2005) and bidirectional siRNA clusters (Kasschau et al., 2007) were removed. Third, computational assessment of foldback structure was done with sequences containing 250 nts on each side of candidate miRNAs using RNAfold, Vienna RNA package, version 1.6.1 (Hofacker, 2003).

miRNA targets were computationally predicted as described (Allen et al., 2005). Briefly, potential targets from FASTA searches (+15/210 match/mismatch scoring ratio, -16 gap penalty and a RNA scoring matrix) were scored using a position-dependent, mispair penalty system (Allen et al., 2005). Penalties were assessed for mismatches, bulges, and gaps (+1 per position) and G:U pairs (+0.5 per position). Penalties were doubled if the mismatch, bulge, gap, or G:U pair occurred at positions 2 to 13 relative to the 5' end of the miRNA. Only one single-nucleotide bulge or single-nucleotide gap was allowed. Based on a reference set of validated miRNA targets, only predicted targets with scores of four or less were considered reasonable. Conservation between *Arabidopsis* and Poplar (*P. trichocarpa*) was assessed by FASTA search, foldback analysis, and detection of similar target sequences (Reinhart et al., 2002; Jones-Rhoades and Bartel, 2004).

miRNA target validation assays

Target validation using a 5'RACE assay was done with the GeneRacer Kit (Invitrogen, CA) as described previously (Llave et al., 2002; Kasschau et al., 2003; Allen et al., 2004; Allen et al., 2005). Poly(A)+ mRNA was isolated from seedling (7 day) and inflorescence tissue (28 day, stage 1–12 flowers) of Col- 0 plants, ligated to adaptor, converted to cDNA and subjected to two rounds of PCR amplification using gene-specific and adaptor-specific primers (Llave et al., 2002; Kasschau et al., 2003; Allen et al., 2004; Allen et al., 2005). Amplified products were gel-purified, cloned and sequenced. Gene-specific primer sequences for miRNA targets that were successfully validated are shown in Table S3.4.

Foldback sequence similarity analysis

The sequences comprising foldbacks from all *MIRNA* loci (Tables 3.1 and 3.2) were identified using RNAfold. Foldback sequences were subjected to FASTA searches against the *Arabidopsis* gene and transcript databases (Allen et al., 2004). The 5' and 3' arms of foldbacks that had gene hits with E-values lower than 0.05 were individually aligned to the top four FASTA hits using NEEDLE (Rice et al., 2000). Each arm was also randomly shuffled 1,000 times using SHUFFLESEQ (Rice et al., 2000) and realigned to each of the top four FASTA hits. The mean ±standard deviation of the randomized sequence scores was calculated. A Z-score was calculated for each arm-gene pair by subtracting the average score of the randomized sequence alignments from the score of the arm-gene alignment and then dividing by the standard deviation of the randomized alignments. This was repeated for arms in which the miRNA or miRNA-complementary sequences were deleted from their respective arms. The deleted arms were aligned with gene sequences in which the target sequence was correspondingly deleted. Intact foldback arms and most-related gene sequences were also aligned and viewed using heat maps with T-COFFEE (Notredame et al., 2000).

ACKNOWLEDGMENTS

We thank Heather Sweet and Amy Shatswell for excellent technical assistance, Christine Gustin and Kathleen Eide for assistance with 5’RACE analysis, and Erin McWhorter for media preparation. We thank Mark Dasenko for sequence analysis. We also thank Detlef Weigel for several interesting discussions and sharing data prior to publication.

Table 3.1. Reference set of *Arabidopsis MIRNA* families

			miRNA*			Target
MIRNA family[a]	Loci	Conserved[b]	sequenced	Reads[c]	Target family	validation[a]
miR156/miR157	12	Y	Y	15606	Squamosa-promoter binding protein-like (SPL)	Y
miR158	2	N	Y	833	Pentatricopeptide repeat (PPR)	N
miR159/miR319	6	Y	Y	24098	MYB transcription factor	Y
					TCP transcription factor	Y
miR160	3	Y	Y	5783	Auxin response factor (ARF)	Y
miR161	1	N	Y	9191	Pentatricopeptide repeat (PPR)	Y
miR162	2	Y	Y	289	Dicer-like (DCL)	Y
miR163	1	N	Y	263	S-adenosyl-methionine-dependent methyltransferase (SAMT)	Y
miR164	3	Y	Y	449	NAC domain transcription factor	Y
miR166/miR165	9	Y	Y	4881	HD-ZIPIII transcription factor	Y
miR167	4	Y	Y	8286	Auxin response factor (ARF)	Y
miR168	2	Y	Y	1286	Argonaute (AGO)	Y
miR169	14	Y	Y	42496	HAP2 transcription factor	Y
miR171/miR170	4	Y	Y	6423	Scarecrow-like (SCL)	Y
miR172	5	Y	Y	2720	Apetala2-like transcription factor (AP2)	Y
miR173	1	N	Y	78	TAS1, TAS2	Y
miR390/miR391	3	Y	Y	2889	TAS3	Y
miR393	2	Y	Y	94	Transport inhibitor response 1 (TIR1)/Auxin F-box (AFB)	Y
					bHLH transcription factor	Y
miR394	2	Y	Y	47	F-box	Y
miR395	6	Y	Y	14	ATP-sulfurylase (APS)	Y
					Sulfate transporter (AST)	Y
miR396	2	Y	Y	408	Growth regulating factor (GRF)	Y
miR397	2	Y	N	290	Laccase (LAC)	Y
miR398	3	Y	Y	42	Copper superoxide dismutase (CSD)	Y
					Cytochrome-c oxidase	Y
miR399	6	Y	Y	50	E2 ubiquiting-conjugating protein (E2-UBC)	Y
miR400	1	N	Y	18	Pentatricopeptide repeat (PPR)	N
miR402	1	N	N	10	HhH-GPD base excision DNA repair	N
miR403	1	Y	Y	29	Argonaute (AGO)	Y
miR408	1	Y	Y	42	Laccase (LAC)	Y
					Plantacyanin-like (PCL)	N
miR447	3	N	N	25	2-phosphoglycerate kinase-related (2-PGK)	Y

[a] Reviewed in Jones-Rhoades et al., (2006).

[b] Conserved between *A. thaliana* and *P. trichocarpa*.

[c] Number of reads are from all libraries in the ASRP database (http://asrp.cgrb.oregonstate.edu/db). Reads for each locus encompass the defined miRNA sequence ±4 nts on each side.

Table 3.2. New or recently identified miRNAs

					Reads[b]			
MIRNA **family**	**Sequence**	**Loci**	**Conserved[a]**	**miRNA* sequenced**	**miRNA**	**miRNA***	**Validated or predicted target family[c]**	**Validated or top predicted targets[c]**
***MIRNA*s with validated targets**								
miR472	UUUUUCCUACUCCGCCCAUACC	1	Y	Y	9 (206)	2 (51)	**CC-NBS-LRR**	**At1g51480 (1)**[d], **At5g43740 (1)**[d], At1g12290 (1.5), [8]
miR773	UUUGCUUCCAGCUUUUGUCUCC	1	N	N	11 (735)	0 (0)	**DNA (cytosine-5-)-methyltransferase**	**At4g14140 (2)**, At4g08990 (3)
miR774	UUGGUUACCCAUAUGGCCAUC	1	N	N	0 (348)	0 (0)	**F-box**	**At3g19890 (1)**[d], At3g17490 (2.5), At3g17265 (3.5)
miR775	UUCGAUGUCUAGCAGUGCCA	1	N	Y	1362 (1532)	5 (3)	**Galactosyltransferase**	**At1g53290 (2.5)**
miR778	UGGCUUGGUUUAUGUACACCG	1	N	Y	0 (2)	1 (7)	**SET-domain**	**At2g22740 (1.5)**, At2g35160 (3.5)[f]
miR780.1	UCUAGCAGCUGUUGAGCAGGU	1	N	Y	57 (266)	0 (118)	**Cation/hydrogen exchanger**	**At5g41610 (3.5)**, At4g33260 (3.5)
miR780.2	UUCUUCGUGAAUAUCUGGCAU							
miR824	UAGACCAUUUGUGAGAAGGGA	1	N	Y	261 (2646)	399 (2)	**MADS-box transcription factor**	**At3g57230 (0.5)**
miR827	UUAGAUGACCAUCAACAAACU	1	N	Y	11 (20)	0 (1)	**SPX domain/Zinc finger (C3HC4-type)**	**At1g02860 (1)**
miR842	UCAUGGUCAGAUCCGUCAUCC	1	N	Y	2 (97)	2 (0)	**Jacalin lectin**	**At5g38550 (2.5)**, At1g60130 (2.5), At1g52120 (2.5), [3]
miR844	AAUGGUAAGAUUGCUUAUAAG	1	N	Y	58 (1)	2 (0)	**Kinase**	**At5g51270 (3.5)[e]**
miR846	UUGAAUUGAAGUGCUUGAAUU	1	N	N	72 (0)	0 (0)	**Jacalin lectin**	**At5g49850 (2.5)[e]**, At5g49870 (2.5)[f], At2g25980 (2.5)[f], [8]
miR856	UAAUCCUACCAAUAACUUCAGC	1	N	Y	62 (7)	9 (0)	**Cation/hydrogen exchanger,** Zinc transporter	**At5g41610 (1)**, At2g46800 (2.5)[f]
miR857	UUUUGUAUGUUGAAGGUGUAU	1	N	N	59 (0)	0 (0)	**Laccase**	**At3g09220 (2)**
miR858	UUUCGUUGUCUGUUCGACCUU	1	N	N	55 (4)	0 (0)	**MYB transcription factor**	**At2g47460 (2.5)**, **At3g08500 (3)**, At5g35550 (3), [6]
miR859	UCUCUCUGUUGUGAAGUCAAA	1	N	N	2 (5)	0 (0)	**F-box**	**At3g17265 (0.5)**[f], At5g36200 (1)[f], **At3g49510 (1.5)**, [32]

Table 3.2. New or recently identified miRNAs (Continued)

MIRNA family	Sequence	Loci	Conserved[a]	miRNA* sequenced	Reads[b]		Validated or predicted target family[c]	Validated or top predicted targets[c]
					miRNA	miRNA*		
***MIRNAs* with only predicted targets, or no predicted targets**								
miR771	UGAGCCUCUGUGGUAGCCCUCA	1	N	Y	16 (906)	0 (44)	-	-
miR776	UCUAAGUCUUCUAUUGAUGUUC	1	N	N	1439 (487)	0 (0)	Serine/threonine kinase	At5g62310 (3)[f]
miR777	UACGCAUUGAGUUUCGUUGCUU	1	N	N	8 (80)	0 (0)	-	-
miR779	UUCUGCUAUGUUGCUGCUCAUU	1	N	N	2 (98)	0 (0)	-	-
miR781	UUAGAGUUUUCUGGAUACUUA	1	N	Y	0 (77)	1 (0)	CD2-binding, MCM	At5g23480 (2.5)[f], At1g44900 (3)
miR823	UGGGUGGUGAUCAUAUAAGAU	1	N	N	107 (1)	0 (0)	Chromomethylase	At1g69770 (2.5)[f]
miR825	UUCUCAAGAAGGUGCAUGAAC	1	N	N	120 (0)	0 (0)	Remorin, zinc finger homeobox family, frataxin-related	At2g41870 (2.5)[f], At5g65410 (3)[f], At4g03240 (3)[f], [1]
miR829.2	AGCUCUGAUACCAAAUGAUGGAAU	1	N	Y	134 (41)	3 (25)	-	-
miR830-5p	UCUUCUCCAAAUAGUUUAGGUU	1	N	Y	2 (1)	-	RanBP1 domain, kinesin motor-related	At1g52380 (3)[f], At3g45850 (3.5)[f]
miR830-3p	UAACUAUUUUGAGAAGAAGUG				-	3 (21)	-	-
miR833-3p	UAGACCGAUGUCAACAAACAAG	1	N	Y	5 (2)	-	-	-
miR833-5p	UGUUUGUUGUACUCGGUCUAG				-	2 (1)	F-box	At1g77650 (3.5)[f]
miR840	ACACUGAAGGACCUAAACUAAC	1	N	Y	20 (1)	7 (116)	WHIRLY transcription factor	At2g02740 (0)[f]
miR843	UUUAGGUCGAGCUUCAUUGGA	1	N	Y	7 (0)	2 (0)	F-box, 1-aminocyclopropane-1-carboxylate synthase	At3g13830 (0.5)[f], At1g11810 (2.5)[f], At2g22810 (3)
miR845a	CGGCUCUGAUACCAAUUGAUG	2	N	Y	670 (21)	1 (40)	-	-
miR845b	UCGCUCUGAUACCAAAUUGAUG							

Table 3.2. New or recently identified miRNAs (Continued)

					Reads[b]			
MIRNA family	Sequence	Loci	Conserved[a]	miRNA* sequenced	miRNA	miRNA*	Validated or predicted target family[c]	Validated or top predicted targets[c]
miR851-5p	UCUCGGUUCGCGAUCCACAAG	1	N	Y	3 (281)	1 (1)	-	-
miR852	AAGAUAAGCGCCUUAGUUCUGA	1	N	Y	3 (84)	0 (1)	ATPase	At5g62670 (3)[f]
miR853	UCCCCUCUUUAGCUUGGAGAAG	1	N	N	2 (0)	0 (0)	-	-
miR860	UCAAUAGAUUGGACUAUGUAU	1	N	Y	14 (15)	0 (1)	Histone deacetylase, ferrochelatase, RNA recognition motif	At5g26040 (0)[f], At5g26030 (0.5)[f], At3g12640 (3.5)
miR861-3p	GAUGGAUAUGUCUUCAAGGAC	1	N	Y	6 (2)	-	-	-
miR861-5p	CCUUGGAGAAAUAUGCGUCAA				-	1 (8)	-	-
miR862-5p	UCCAAUAGGUCGAGCAUGUGC	1	N	Y	5 (0)	-	-	-
miR862-3p	AUAUGCUGGAUCUACUUGAAG				-	2 (0)	-	-
miR863-3p	UUGAGAGCAACAAGACAUAAU	1	N	Y	5 (0)	-	-	-
miR863-5p	UUAUGUCUUGUUGAUCUCAAU				-	2 (0)	Kinase, Legumain (C13 protease)	At2g26700 (3), At1g62710 (3.5)[f]
miR864-5p	UCAGGUAUGAUUGACUUCAAA	1	N	Y	3 (0)	-	Triacylglycerol lipase	At1g06250 (3)[f]
miR864-3p	UAAAGUCAAUAAUACCUUGAAG				-	2 (0)	Expressed protein	At4g25210 (3)[f]
miR865-5p	AUGAAUUUGGAUCUAAUUGAG	1	N	Y	3 (0)	-	Serine carboxypeptidase, sulfate transporter	At5g42240 (3.5)[f], At3g51895 (3.5)[f]
miR865-3p	UUUUUCCUCAAAUUUAUCCAA				-	1 (0)	DEAD box RNA helicase, DNA-binding bromodomain-containing protein	At2g07750 (3)[f], At2g34900 (3)[f], At1g03770 (3.5), [6]
miR866-3p	ACAAAAUCCGUCUUUGAAGA	1	N	Y	2 (0)	-	Kinase, electron transport SCO1/SenC, NAD-dependent G-3-P dehydrogenase	At4g21400 (3)[f], At4g39740 (3), At2g41540 (3)[f], [4]
miR866-5p	UCAAGGAACGGAUUUUGUUAA[g]				-	0 (5)	Expressed protein, C2 domain-containing protein	At4g21700 (3)[f], At1g66360 (3.5)[f]

Table 3.2. New or recently identified miRNAs (Continued)

					Reads[b]			
MIRNA family	Sequence	Loci	Conserved[a]	miRNA* sequenced	miRNA	miRNA*	Validated or predicted target family[c]	Validated or top predicted targets[c]
miR867	UUGAACAUGGUUUAUUAGGAA	1	N	N	30 (0)	0 (0)	PHD finger-related/SET domain, kinase, phospholipase/carboxyl-esterase	At4g27910 (3.5)[f], At3g17750 (3.5)[f], At3g15650 (3.5)[f]
miR868	CUUCUUAAGUGCUGAUAAUGC	1	N	Y	9 (1)	0 (0)	-	-
miR869.1	UCUGGUGUUGAGAUAGUUGAC	1	N	N	11 (5)	0 (0)	-	-
miR869.2	AUUGGUUCAAUUCUGGUGUUG							
miR870	UAAUUUGGUGUUUCUUCGAUC	1	N	N	4 (32)	0 (0)	-	-

[a] Conserved between *A. thaliana* and *P. trichocarpa*.

[b] Number of reads are from all libraries in the ASRP (http://asrp.cgrb.oregonstate.edu/db) and MPSS Plus (http://mpss.udel.edu/at/) databases. Reads for each locus encompass the defined miRNA/miRNA* sequence ±4 nts on each side. ASRP (MPSS Plus).

[c] Top three predicted targets with a score of 3.5 or less are listed with their score in parentheses. Targets validated by 5'RACE are in bold. Remaining number of targets predicted with a score of 3.5 or less are listed in square brackets (Table S3.1). Dashes indicate no predicted targets with a score of 3.5 or less.

[d] Targets validated by Lu et al. (2006).

[e] Target tested but failed in 5'RACE validation assays.

[f] Seventeen nt MPSS Plus signature was extended 4 nts on the 3' end.

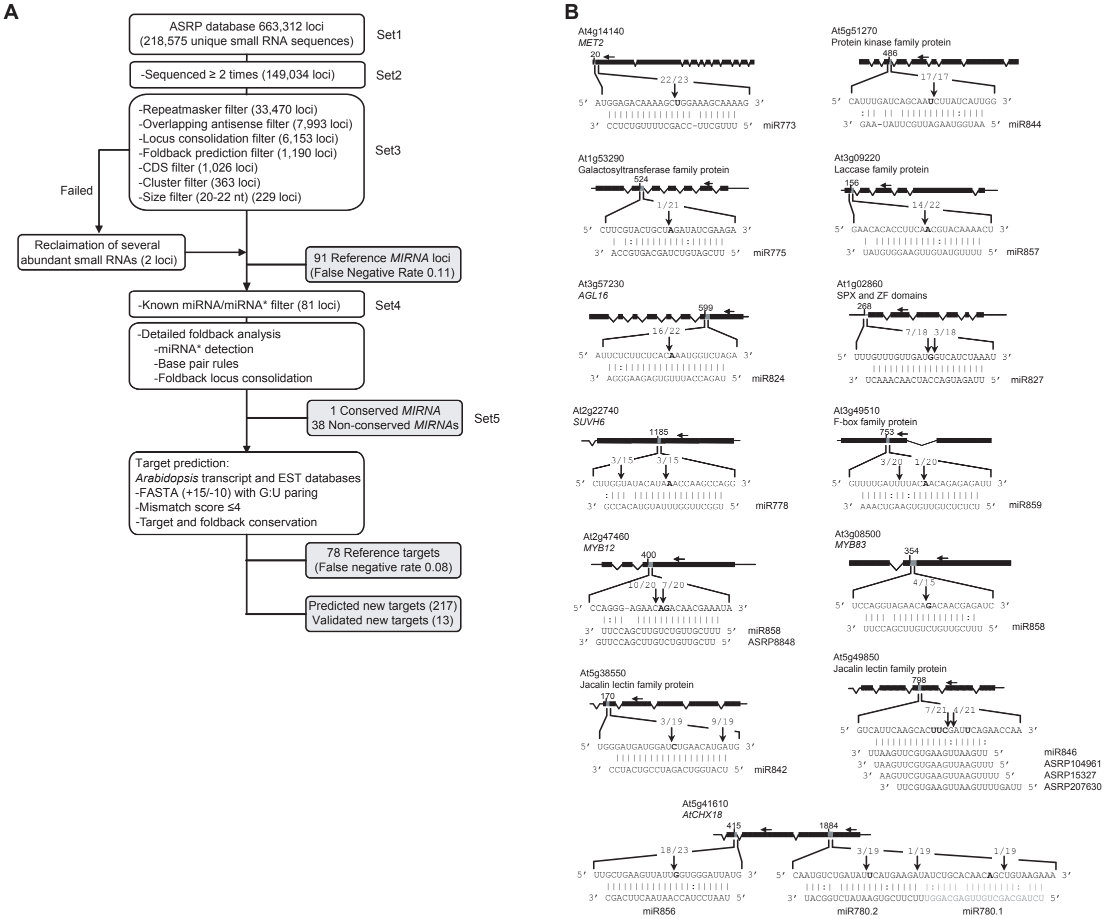

A
ASRP database 663,312 loci
(218,575 unique small RNA sequences)
Set1
-Sequenced ≥ 2 times (149,034 loci)
Set2
-Repeatmasker filter (33,470 loci)
-Overlapping antisense filter (7,993 loci)
-Locus consolidation filter (6,153 loci)
-Foldback prediction filter (1,190 loci)
-CDS filter (1,026 loci)
-Cluster filter (363 loci)
-Size filter (20-22 nt) (229 loci)
Set3
Failed
Reclaimation of several
abundant small RNAs (2 loci)
91 Reference MIRNA loci
(False Negative Rate 0.11)
-Known miRNA/miRNA* filter (81 loci)
Set4
-Detailed foldback analysis
-miRNA* detection
-Base pair rules
-Foldback locus consolidation
1 Conserved MIRNA
38 Non-conserved MIRNAs
Set5
Target prediction:
Arabidopsis transcript and EST databases
-FASTA (+15/-10) with G:U paring
-Mismatch score ≤4
-Target and foldback conservation
78 Reference targets
(False negative rate 0.08)
Predicted new targets (217)
Validated new targets (13)
B
At4g14140
MET2
20
22/23
miR773
At5g51270
Protein kinase family protein
486
17/17
miR844
At1g53290
Galactosyltransferase family protein
524
1/21
miR775
At3g09220
Laccase family protein
156
14/22
miR857
At3g57230
AGL16
599
16/22
miR824
At1g02860
SPX and ZF domains
268
7/18 3/18
miR827
At2g22740
SUVH6
1185
3/15 3/15
miR778
At3g49510
F-box family protein
753
3/20 1/20
miR859
At2g47460
MYB12
400
10/20 7/20
miR858
ASRP8848
At3g08500
MYB83
354
4/15
miR858
At5g38550
Jacalin lectin family protein
170
3/19 9/19
miR842
At5g49850
Jacalin lectin family protein
798
7/21 4/21
miR846
ASRP104961
ASRP15327
ASRP207630
At5g41610
AtCHX18
415
1884
18/23
3/19
1/19
1/19
miR856
miR780.2
miR780.1

Figure 3.1. Identification and analysis of *Arabidopsis* miRNAs and targets.
(A) Flowchart for the prediction of miRNAs and their targets. **(B)** Validation of predicted targets for 13 non-conserved miRNAs. Positions of dominant 5'RACE products (no. 5' ends at position/total no. 5' ends sequenced) are indicated by vertical arrows in the expanded regions. Predicted cleavage sites are indicated by a bolded nucleotide at position ten relative to the 5' end of the miRNA or miRNA-variant. Positions of gene-specific primers are indicated with horizontal arrows above gene diagrams.

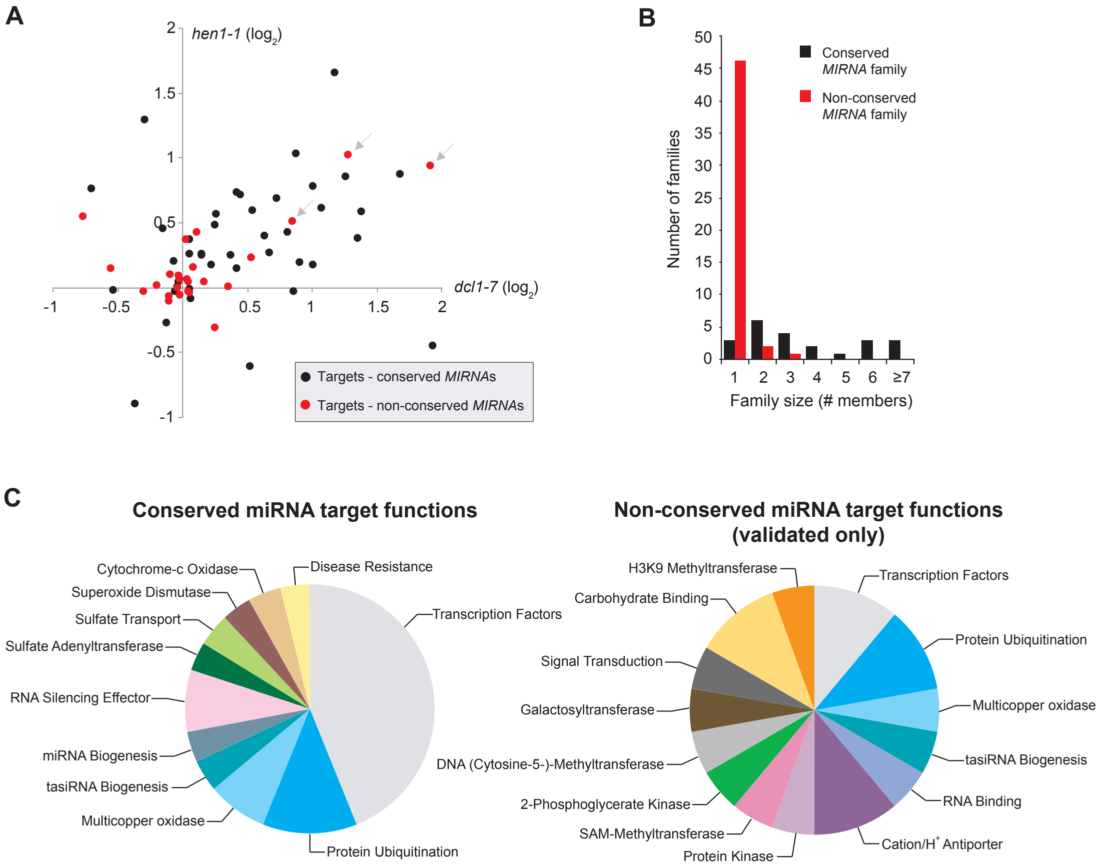


Figure 3.2. Comparison of conserved and non-conserved *MIRNA* families. **(A)** Effect of *dcl1-7* and *hen1-1* mutations on levels of target transcripts for conserved (black) and non-conserved (red) miRNAs. Expression data are shown for two validated or high-confidence predicted targets, if available, for each family. Arrows indicate targets for miR824 (*AGL16*), miR858 (*MYB12*) and miR161.1 (At1g63130, a *PPR* gene). **(B)** Numbers of gene family members for conserved and non-conserved *MIRNA*s (Tables 3.1 and 3.2). **(C)** Relative numbers of miRNA target family functions for conserved and non-conserved miRNAs (Tables 3.1 and 3.2). Only target classes that have been validated experimentally are included. Note that Table 3.2 shows many *MIRNA* families with weak or no predicted targets, and these are not represented in the chart.

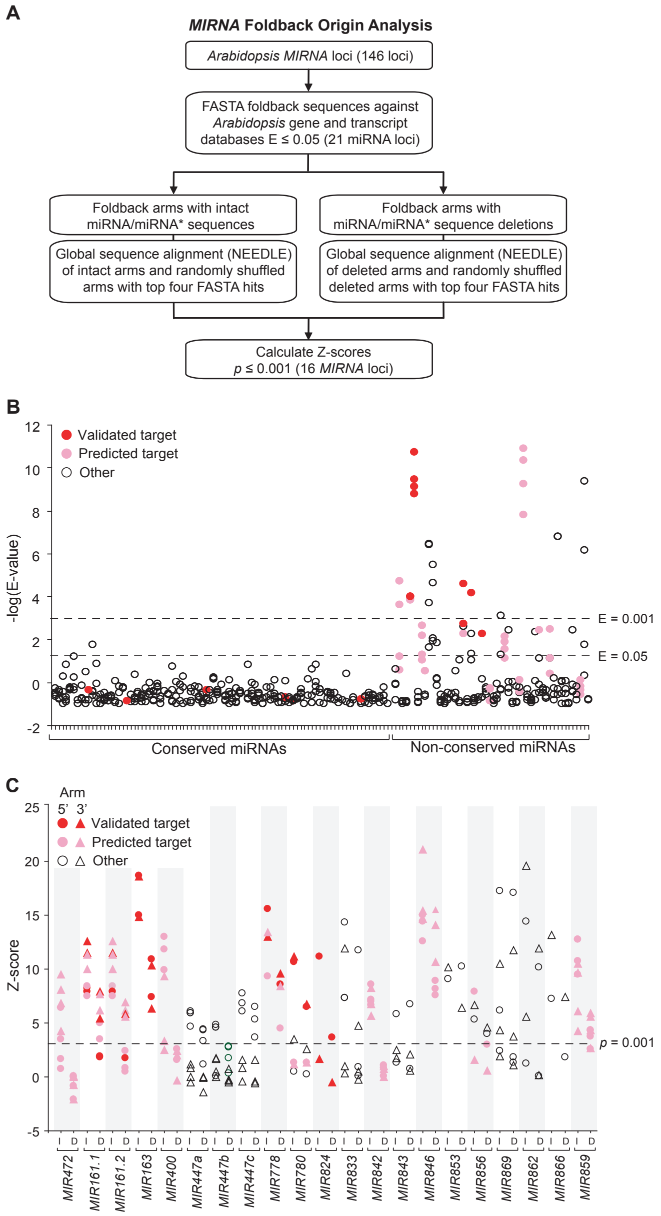


Figure 3.3. Identification of *MIRNA* foldbacks with similarity to protein-coding genes. **(A)** Flowchart for identification of *MIRNA* foldbacks with similarity, extending beyond the miRNA target site, to protein-coding genes. **(B)** *Arabidopsis* gene or transcript hits in FASTA searches using foldback sequences for all conserved and non-conserved *MIRNA* loci (Tables 3.1 and 3.2). The top four hits based on E-values are shown. **(C)** Z-scores for the Needleman-Wunsch alignment values from *MIRNA* foldback arms with top four gene or transcript FASTA hits. Alignments were done with intact foldback arms (I), and with foldback arms in which miRNA or miRNA-complementary sequences were deleted (D). Z-scores were derived from standard deviation values for alignments of randomized sequences. In **(B)** and **(C)**, a red symbol represents an experimentally validated target, a pink symbol indicates a gene from a validated target family, and an open symbol indicates a gene that is distinct from either the validated or predicted target family.

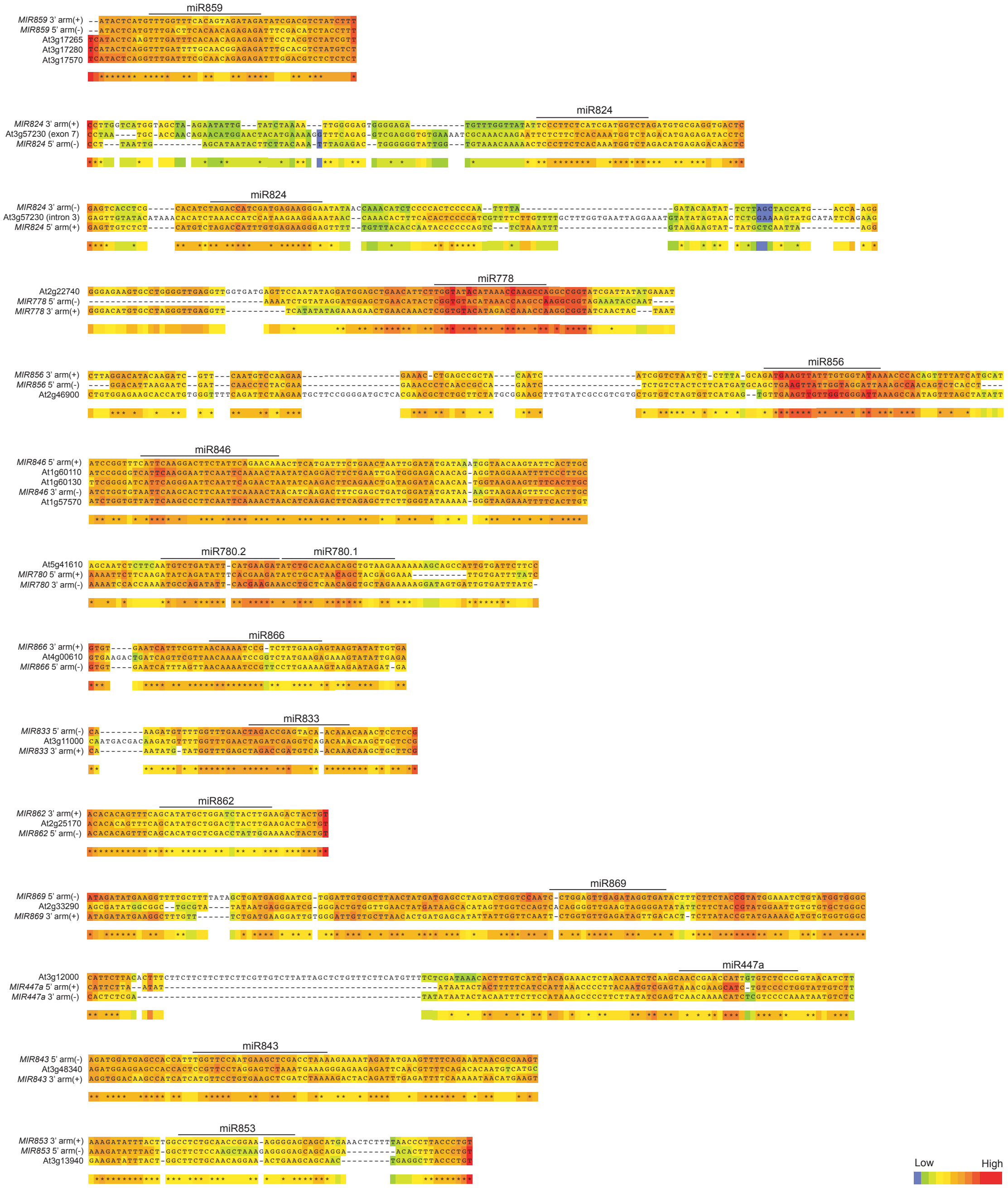


Figure 3.4. Similarity between *MIRNA* foldback arms and protein-coding genes. Each alignment contains the coding strand for 1–3 genes, the miRNA* arm, and the miRNA arm. Orientation of the foldback arms is indicated by (+) for authentic polarity and (–) for the reverse complement polarity. Two alignments are given for *MIR824* because the two arms are each most similar to distinct, duplicated regions within the *AGL16* gene (At3g57230). Alignments were generated using T-COFFEE. Colors indicate alignment quality in a regional context.

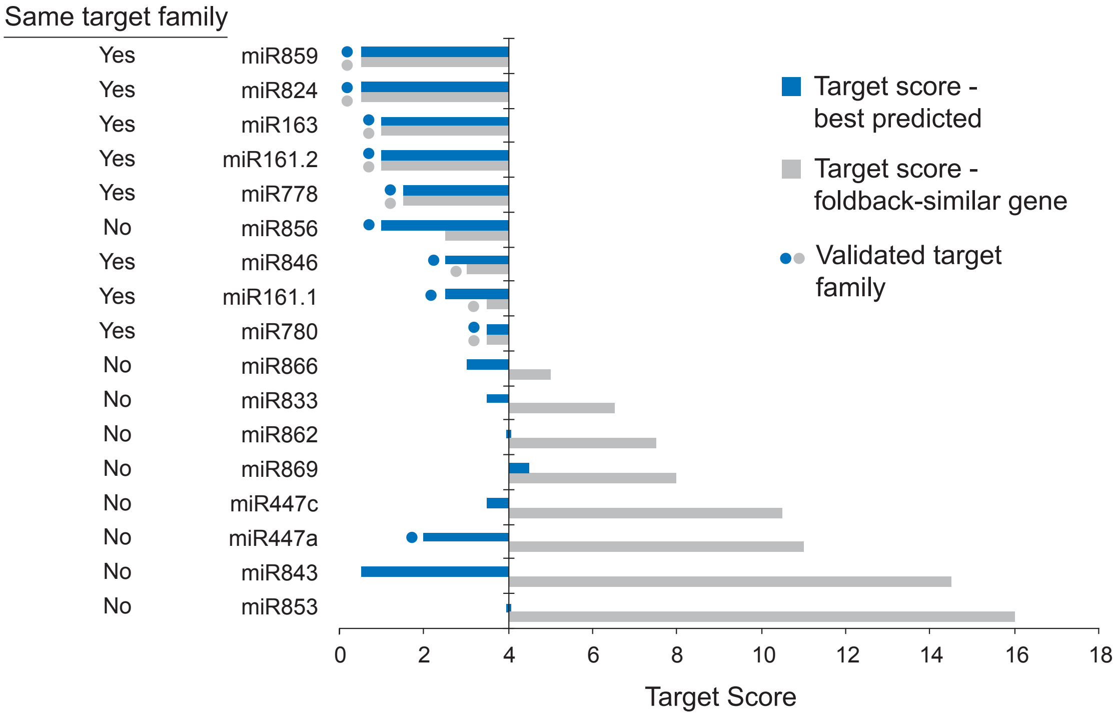


Figure 3.5. Targeting specificity of recently evolved *MIRNA*s.
Two target prediction scores are shown for each of 16 miRNAs: best overall predicted target score (blue) and target scores calculated for *MIRNA* foldback-similar genes (grey). Left column indicates whether or not the best overall predicted target gene is in the same family as the foldback-similar gene. A dot indicates that the predicted gene is in an experimentally validated target family. Two calculations corresponding to the two major populations from the *MIR161* locus (miR161.1 and miR161.2) are shown. The identities of targets are listed in Table S3.3. The plot is centered on a target prediction score of 4, as this corresponds to the upper limit of a reasonable prediction.

SUPPLEMENTAL DATA

Table S3.1. Additional predicted targets for eight *MIRNA*s

***MIRNA* family**	**Predicted targets[a]**
miR472	At1g12210 (2.5), At5g63020 (2.5), At1g12220 (3), At1g15890 (3), At4g10780 (3), At5g43730 (3), At1g12280 (3.5), At5g47260 (3.5)
miR825	At5g44970 (3.5)
miR842	At1g52100 (3.5), At5g28520 (3.5), At1g52070 (3.5)
miR846	At1g57570 (3)[b], At1g52130 (3), At1g33790 (3), At1g60110 (3.5), At5g28520 (3.5), At1g52070 (3.5), At1g52050 (3.5), At1g52060 (3.5)
miR858	At1g06180 (3), At5g49330 (3), At1g66230 (3), At4g12350 (3), At3g62610 (3.5), At3g24310 (3.5)
miR859	At2g18780 (1.5), At5g36820 (1.5), At5g36730 (1.5), At3g17570 (1.5), At3g22350 (1.5), At3g16880 (1.5), At3g22710 (1.5), At3g16820 (1.5), At3g21170 (2.5), At3g49520 (2.5), At3g24580 (2.5), At2g27520 (2.5), At3g13820 (2.5), At3g13830 (3), At3g22700 (3), At1g11810 (3), At3g19880 (3.5), At3g17280 (3.5), At2g04920 (3.5), At1g25054 (3.5), At1g25141 (3.5), At1g25210 (3.5), At1g24793 (3.5), At1g24880 (3.5), At1g32140 (3.5), At4g33290 (3.5), At1g67450 (3.5), At3g17540 (3.5), At3g22720 (3.5), At3g20710 (3.5), At2g24510 (3.5), At5g42460 (3.5)
miR865-3p	At3g02340 (3.5), At3g10160 (3.5),At1g05680 (3.5), At1g31290 (3.5), At1g64880 (3.5), At5g67450 (3.5)
miR866-3p	At1g09680 (3), At2g22070 (3.5), At3g05380 (3.5), At5g18370 (3.5)

[a] Remaining predicted targets from Table 3.2 with a score of 3.5 or less. Target score is given in parentheses.

[b] Target tested but not validated by 5'RACE.

Table S3.2. Summary of *MIRNA* foldback sequences

MIRNA family	Location of foldback[a]	Location of small RNA[a]	Strand	ΔG (kcal/mol)
miR472	1:4182133..4182294	1:4182141..4182162	-1	-54.80
miR771	3:19670274..19670397	3:19670359..19670380	-1	-55.40
miR773	1:13067199..13067342	1:13067291..13067312	1	-61.20
miR774	1:22153601..22153698	1:22153670..22153690	1	-40.40
miR775	1:29427347..29427465	1:29427439..29427458	1	-41.33
miR776	1:22799293..22799391	1:22799364..22799385	1	-41.70
miR777	1:26641663..26641770	1:26641746..26641767	1	-39.80
miR778	2:17356515..17356652	2:17356626..17356646	-1	-79.70
miR779	2:9567831..9568011	2:9567982..9568003	1	-71.60
miR780.1	4:8504127..8504318	4:8504161..8504181	-1	-71.60
miR780.2		4:8504140..8504160		
miR781	1:7423507..7423600	1:7423512..7423532	1	-42.00
miR823	3:4496829..4496925	3:4496833..4496853	-1	-42.30
miR824	4:12625121..12625766	4:12625138..12625158	1	-187.02
miR825	2:11166780..11166881	2:11166852..11166872	1	-42.10
miR827	3:22133773..22133876	3:22133788..22133808	-1	-35.30
miR829.1	1:11834048..11834160	1:11834074..11834097	-1	-60.50
miR830-5p	1:4820402..4820496	1:4820470..4820491	-1	-26.25
miR830-3p		1:4820403..4820423		
miR833-3p	1:29530087..29530187	1:29530158..29530179	1	-57.50
miR833-5p		1:29530097..29530117		
miR840	2:771376..771521	2:771491..771512	-1	-74.40
miR842	1:22580778..22580884	1:22580859..22580879	1	-51.50
miR843	3:17753094..17753276	3:17753132..17753152	1	-87.10
miR844	2:9949273..9949373	2:9949343..9949363	-1	-47.40
miR845a	4:12217457..12217537	4:12217467..12217487	-1	-42.40
miR845b	4:12214079..12214177	4:12214096..12214117	-1	-41.10
miR846	1:22581082..22581356	1:22581327..22581347	1	-94.70
miR851-5p	3:19670549..19670654	3:19670624..19670644	-1	-56.20
miR852	4:8336203..8336327	4:8336298..8336319	1	-65.40
miR853	3:8346400..8346646	3:8346423..8346444	1	-85.20
miR856	1:11957440..11957710	1:11957467..11957488	1	-110.60
miR857	4:7878181..7878557	4:7878194..7878214	-1	-116.86
miR858	1:26777201..26777387	1:26777357..26777377	-1	-49.34
miR859	1:22153383..22153508	1:22153400..22153420	1	-63.42
miR860	5:9098794..9098919	5:9098879..9098899	1	-64.70
miR861-3p	3:17849120..17849251	3:17849139..17849159	-1	-56.30
miR861-5p		3:17849208..17849228		
miR862-5p	2:10725175..10725277	2:10725185..10725205	1	-62.70
miR862-3p		2:10725249..10725269		
miR863-3p	4:7846593..7846895	4:7846831..7846851	1	-140.20
miR863-5p		4:7846640..7846660		
miR864-5p	1:6740491..6740582	1:6740494..6740514	1	-26.27
miR864-3p		1:6740560..6740581		
miR865-5p	5:5169993..5170134	5:5170001..5170021	1	-43.00
miR865-3p		5:5170100..5170120		
miR866-3p	5:16175051..16175176	5:16175066..16175085	-1	-58.60
miR866-5p		5:16175159..16175139		
miR867	4:11375375..11375492	4:11375398..11375418	1	-39.10
miR868	3:6488240..6488432	3:6488405..6488425	1	-83.70
miR869.1	5:15908807..15909101	5:15908842..15908862	-1	-163.10
miR869.2		5:15908853..15908873		
miR870	5:21412771..21412855	5:21412771..21412791	-1	-42.20

[a] Coordinates are listed as chromosome:start nucleotide..end nucleotide.

Table S3.3. *MIRNA* loci with sequence similarity to protein-coding genes

	Foldback-similar genes			Best predicted target genes		
miRNA	Gene[a]	Gene family	Target score	Gene[a]	Gene family	Target score
miR161.1[b]	**At5g41170**	Pentatricopeptide repeat	3.5	**At1g63400**	Pentatricopeptide repeat	2.5
				At1g63130	Pentatricopeptide repeat	2.5
				At1g63080	Pentatricopeptide repeat	2.5
				At1g62910[c]	Pentatricopeptide repeat	2.5
				At1g64580	Pentatricopeptide repeat	2.5
				At1g62670	Pentatricopeptide repeat	2.5
				At1g62930	Pentatricopeptide repeat	2.5
miR161.2[b]	At5g41170	Pentatricopeptide repeat	1	At5g41170	Pentatricopeptide repeat	1
				At1g62590	Pentatricopeptide repeat	1
				At1g63330	Pentatricopeptide repeat	1
miR163[b]	**At1g66690**	S-adenosyl-L-methionine:carboxyl methyltransferase	1	**At1g66690**	S-adenosyl-L-methionine:carboxyl methyltransferase	1
				At1g66700	S-adenosyl-L-methionine:carboxyl methyltransferase	1
miR447a[b]	At3g12000	S-locus related	11	**At5g60760**	2-phosphoglycerate kinase-related	2
miR447c[b]	At4g21370	S-locus protein kinase	10.5	At5g42870	Lipin	3.5
miR778[b]	**At2g22740** (SUVH6)	H3K9 methyltransferase	1.5	**At2g22740** (SUVH6)	H3K9 methyltransferase	1.5
miR780	**At5g41610** (CHX18)	Sodium hydrogen antiporter	3.5	**At5g41610** (CHX18)	Sodium hydrogen antiporter	3.5
miR824	**At3g57230** (AGL16)	MADS-box	0.5	**At3g57230** (AGL16)	MADS-box	0.5
miR833-5p	At3g11000	Development and cell death domain	6.5	At1g77650[c]	F-box	3.5
miR843	At3g48340	Cysteine proteinase	14.5	At3g13830[c]	F-box	0.5
miR846[b]	At1g57570[c]	Jacalin lectin	3	At2g25980[c]	Jacalin lectin	2.5
				At5g49850	Jacalin lectin	2.5
				At5g49870[c]	Jacalin lectin	2.5
miR853	At3g13940	A49-like RNA Polymerase I associated factor	16	At5g08010[c]	Unclassified	4
miR856	At2g46800 (ZAT1)[c]	Zinc transporter	2.5	**At5g41610** (CHX18)	Sodium hydrogen antiporter	1
miR859[b]	At3g17265[c]	F-box	0.5	At3g17265[c]	F-box	0.5
miR862	At2g25170 (PKL)	Chromatin remodeling factor	7.5	At5g11530 (EMF1)[c]	Embryonic flower 1	4
miR866-5p	At4g00610	DNA-binding storekeeper	5	At4g21700[c]	Unclassified	3
miR869[bd]	At2g33290 (SUVH2)	H3K9 methyltransferase	8	At1g22400 (UGT85A1)	UDP-glucoronosyl/UDP-glucosyl transferase	4.5
				At1g71230 (AJH2)	COP9 signalosome subunit 5A	4.5

[a] Genes in bold were validated for miRNA-guided cleavage by 5'RACE.

[b] *MIRNA locus* has significant similarity to multiple genes in the same family.

[c] Target was tested but failed the validation assay.

[d] Targets predicted for miRNA*

Table S3.4. Gene-specific PCR primers for successful validation of miRNA targets by 5' RACE

miRNA	Target gene	PCR round	Primer name	Primer sequence (5' to 3')
miR856	At5g41610	1st Rd	At5g41610_miR856_512	GTGGAATGATGAAAGACGCACCGATA
		2nd Rd	At5g41610_miR856_406	AGTAGAATCCAAGCTGCCACATCGT
miR824	At3g57230	1st Rd	At3g57230_miR824_353	CAACGGAAGGGTGACAACTTTATGCT
		2nd Rd	At3g57230_miR824_236	GCTGGTGTTTGATAGATATGGAATGCA
miR775	At1g53290	1st Rd	At1g53290_miR775_419	GAAGGTGAGCATTCTGGTTCGCAAA
		2nd Rd	At1g53290_miR775_419	GAAGGTGAGCATTCTGGTTCGCAAA
miR858	At3g08500	1st Rd	At3g08500_miR858_412	TGAAGCAAGGATCAAGGGCCTGTAA
		2nd Rd	At3g08500_miR858_412	TGAAGCAAGGATCAAGGGCCTGTAA
miR773	At4g14140	1st Rd	At4g14140_miR773_452	TCGTAATCCGCTAACGCCGTTGAAAT
		2nd Rd	At4g14140_miR773_417	CGGGGAACCATCTTCATAACCAGATA
miR844	At5g51270	1st Rd	At5g51270_miR844_551	GCGTGTCCTGGTTGTCTGTGAAATA
		2nd Rd	At5g51270_miR844_258	GGGTTAGATTTGAGTGCATTCGACATA
miR857	At3g09220	1st Rd	At3g09220_miR857_494	GCCATTGATCGTATAAGCGTCGGAATT
		2nd Rd	At3g09220_miR857_368	GGGAAACGGGTAAGAGTGACCAGATT
miR827	At1g02860	1st Rd	At1g02860_miR827_377	GCCTCTGAGCCAAGTAAGACACTTT
		2nd Rd	At1g02860_miR827_377	GCCTCTGAGCCAAGTAAGACACTTT
miR858	At2g47460	1st Rd	At2g47460_miR858_506	CCTCTGGCTCCTTCAGAGTCTCTTA
		2nd Rd	At2g47460_miR858_431	CCAGGATCTGACTCGTCCAACAAAA
miR842	At5g38550	1st Rd	At5g38550_miR842_432	CATCGTCCCACTTCTTGCTTCCCTTA
		2nd Rd	At5g38550_miR842_216	GCATACCGTAGGACTTTTTGTCGTAGT
miR780	At5g41610	1st Rd	At5g41610_miR780_424	CTCCTCATCCGATTTCAGATTCTTCAC
		2nd Rd	At5g41610_miR780_529	TTCCTCAATCGCCGATCTAACATCCA
miR846	At5g49850	1st Rd	At5g49850_miR846_330	CGTCGTTCATTCCATGGGCACGCT
		2nd Rd	At5g49850_miR846_478	GAAGTTCTCTCACCAAATGTCCTAGAG
miR859	At3g49510	1st Rd	At3g49510_miR859_275	CCTCGTCAATGAAGAAGCTCTTAGCT
		2nd Rd	At3g49510_miR859_357	CCATCATCTCCAACGATGTAAGCCAT
miR778	At2g22740	1st Rd	At2g22740_miR778_225	CTATGCTGGTTGCTAACGCGAGGTT
		2nd Rd	At2g22740_miR778_353	CTGAGATCCGACTTGTTGCCAATAC

MicroRNA Gene Evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*

Noah Fahlgren, Sanjuro Jogdeo, Kristin D. Kasschau, Christopher M. Sullivan, Elisabeth J. Chapman, Sascha Laubinger, Lisa M. Smith, Mark Dasenko, Scott A. Givan, Detlef Weigel and James C. Carrington

The Plant Cell, American Society of Plant Biologists
15501 Monona Drive
Rockville, Maryland 20855-2768
Volume 22, pages 1074-1089
Reprinted with permission

SUMMARY

MicroRNAs (miRNAs) are short regulatory RNAs processed from partially self-complementary foldbacks within longer *MIRNA* primary transcripts. Several *MIRNA* families are conserved deeply through land plants, but many are present only in closely related species or are species specific. The finding of numerous evolutionarily young *MIRNA*, many with low expression and few if any targets, supports a rapid birth-death model for *MIRNA* evolution. A systematic analysis of *MIRNA* genes and families in the close relatives, *Arabidopsis thaliana* and *Arabidopsis lyrata*, was conducted using both whole-genome comparisons and high-throughput sequencing of small RNAs. Orthologs of 143 *A. thaliana MIRNA* genes were identified in *A. lyrata*, with nine having significant sequence or processing changes that likely alter function. In addition, at least 13% of *MIRNA* genes in each species are unique, despite their relatively recent speciation (~10 million years ago). Alignment of *MIRNA* foldbacks to the *Arabidopsis* genomes revealed evidence for recent origins of 32 families by inverted or direct duplication of mostly protein-coding gene sequences, but less than half of these yield miRNA that are predicted to target transcripts from the originating gene family. miRNA nucleotide divergence between *A. lyrata* and *A. thaliana* orthologs was higher for young *MIRNA* genes, consistent with reduced purifying selection compared with deeply conserved *MIRNA* genes. Additionally, target sites of younger miRNA were lost more frequently than for deeply conserved families. In summary, our systematic analyses emphasize the dynamic nature of the *MIRNA* complement of plant genomes.

INTRODUCTION

MicroRNA (miRNA) are a class of small RNA encoded in the genomes of plants, animals, algae, some other unicellular organisms, and many DNA viruses (Carthew and Sontheimer, 2009; Cullen, 2009; Voinnet, 2009). Primary transcripts from *MIRNA* genes form imperfect stem-loop structures that are processed by one (plants) or two (animals) RNaseIII domain nucleases in the Dicer family, which function with accessory RNA binding proteins and components of the nuclear cap binding complex (Carthew and Sontheimer, 2009; Voinnet, 2009). The resulting miRNA-miRNA* duplexes undergo 2'-O-methylation, and the miRNA strand associates with a member of the Argonaute (AGO) protein family through several specificity mechanisms (Carthew and Sontheimer, 2009; Voinnet, 2009). miRNA-AGO complexes interact with miRNA complementary sites within target transcripts, usually with the effect of target transcript repression through degradative or nondegradative mechanisms (Carthew and Sontheimer, 2009; Voinnet, 2009).

Although the biogenesis and effector mechanisms for eukaryotic miRNA involve factors that originated in ancient eukaryotes (Shabalina and Koonin, 2008), there are no convincing examples of miRNA conserved between plants and animals, suggesting that *MIRNA* genes evolved independently in plants and animals (Axtell, 2008). Several mechanisms for forming new *MIRNA* genes have been proposed. In plants, evidence of extensive sequence similarity between foldback sequences and protein-coding loci was found for several young *Arabidopsis MIRNA* genes, suggesting that *MIRNA* can form by inverted duplication events (Allen et al., 2004; Rajagopalan et al., 2006; Axtell et al., 2007; Fahlgren et al., 2007). Initially, these *MIRNA* would have a high degree of complementarity to the parental locus and, if expressed, could produce small RNA that target the parental transcript. In animals, no evidence of inverted duplication-driven *MIRNA* formation has been found (Chen and Rajewsky, 2007). Rather, unique animal *MIRNA* may originate from numerous hairpins in the genome by chance acquisition of expression and miRNA-processing characteristics (Chen and Rajewsky, 2007). Evidence for spontaneous formation of *MIRNA* genes was reported in *Drosophila* species (Lu et al., 2008b) and has also been proposed for some *Arabidopsis MIRNA* (de Felippes et al., 2008). Additionally, transposable elements may have been the source of some animal *MIRNA* (Smalheiser and Torvik, 2005; Borchert et al., 2006; Piriyapongsa and Jordan, 2007; Piriyapongsa et al., 2007) and potentially some plant *MIRNA* (Piriyapongsa and Jordan, 2008). The inverted repeats found in some classes of transposable elements could be the raw material for hairpin RNA that, if processed, might generate small RNA that target similar repetitive sequences integrated into transcribed genes (Smalheiser and Torvik, 2005; Piriyapongsa and Jordan, 2007; Piriyapongsa et al., 2007).

In both plants and animals, some *MIRNA* families are highly conserved through hundreds of millions of years (Axtell and Bartel, 2005; Grimson et al., 2008). However, individual species also contain highly specific, recently evolved *MIRNA* genes (Chen and Rajewsky, 2007; Voinnet, 2009). Deeply conserved *MIRNA* families have expanded and specialized by duplication and sub- or neofunctionalization (Maher et al., 2006; Chen and Rajewsky, 2007; Rubio-Somoza et al., 2009), whereas young *MIRNA* may initially evolve neutrally (Chen and Rajewsky, 2007; Axtell, 2008). The appearance of large numbers of relatively young *MIRNA* suggests that lineage-specific *MIRNA* are born frequently but are also lost frequently (Rajagopalan et al., 2006; Fahlgren et al., 2007; Axtell, 2008; Lu et al., 2008b). In plants, the exact frequency of births and deaths has not been determined, since so far only genome sequences from relatively distantly related species have been available for comparison.

In *Arabidopsis thaliana*, loci encoding perfect or near-perfect hairpins that yield heterogeneous small RNA through the activity of multiple DICER-LIKE (DCL) proteins were proposed to be the evolutionary precursors of canonical *MIRNA* genes. This idea is supported by the fact that two young *A. thaliana MIRNA*, *ath-MIR822* and *ath-MIR839*, produce heterogeneous small RNA that are dependent on the activity of DCL4 (Rajagopalan et al., 2006). Similarly, long hairpin RNA in *Drosophila* are processed heterogeneously due to crosstalk between factors involved in the miRNA and small interfering (siRNA) pathways (Okamura et al., 2008). Over time, substitutions reducing self-complementarity in the hairpin may decrease the efficiency of these hairpins entering siRNA-generating pathways, while subjecting them to processing by the miRNA biogenesis machinery (Chapman and Carrington, 2007).

Young miRNAs that arose through the inverted duplication mechanism could potentially be deleterious because of suppressive interactions with transcripts from the originating gene family. These would presumably be lost through purifying selection. If expressed at low levels or in a restricted manner, deleterious effects might be minimized. In such cases, the evolutionary window during which neutral substitutions can accumulate should be longer, and such loci should be more easily found (Chen and Rajewsky, 2007). In fact, deeply conserved plant *MIRNA* families tend to be expressed more abundantly than younger *MIRNA* (Lu et al., 2006; Rajagopalan et al., 2006; Fahlgren et al., 2007; Axtell, 2008), and human miRNAs that are lowly expressed are evolving neutrally (Liang and Li, 2009). In plants, compared with deeply conserved miRNAs, young miRNAs have been associated with fewer target transcripts (Rajagopalan et al., 2006; Axtell et al., 2007; Fahlgren et al., 2007). Given the low expression levels, the high proportion that lack targets, and the evidence for high birth and death rates, most lineage-specific *MIRNA* may be evolutionarily transient loci that are evolving neutrally (Axtell, 2008). However, in rare cases, target interactions could be formed and fixed in a population, leading to maintenance of the *MIRNA* locus. For example, the relatively young miR824 functions within a leaf patterning regulatory network by targeting *AGOMOUS-LIKE16*, a member of the *MIR824*-originating MADS box family (Kutter et al., 2007). In other cases, mutations may cause targeting to shift to transcripts unrelated to the locus that gave rise to the *MIRNA* (Fahlgren et al., 2007; de Felippes et al., 2008).

The recent determination of the genome sequence of *Arabidopsis lyrata* (http://genome.jgi-psf.org/Araly1/Araly1.home.html), a species that diverged from *A. thaliana* ~10 million years ago (Koch et al., 2000; Wright et al., 2002; Ossowski et al., 2010), provides an opportunity to assess evolutionary histories for the many *MIRNA* found previously only in *A. thaliana*. Here, we describe the genome-wide small RNA landscape of *A. lyrata* and identify *MIRNA* shared between *A. thaliana* and *A. lyrata*, as well as *MIRNA* that are not shared between the two species. We reinvestigate the origins of *MIRNA* loci and find additional evidence for duplication-type origins from both coding and noncoding loci and from a repetitive element. Furthermore, we provide evidence supporting the idea that many young *MIRNA* are evolving neutrally and are found in genomic regions in a higher state of flux. Finally, we report that interactions between young *MIRNA* and targets are highly fluid relative to those involving deeply conserved *MIRNA* families.

RESULTS

A. lyrata small RNA landscape

The recently completed genome sequence of *A. lyrata* (http:// genome.jgi-psf.org/Araly1/Araly1.home.html), along with the established sequence of *A. thaliana* (Arabidopsis Genome Initiative, 2000), provides the opportunity to compare RNA silencing systems of two closely related plant species. Small RNA libraries were constructed for *A. lyrata* and analyzed initially by high-throughput pyrosequencing (454 Life Sciences) and then using sequencing-by-synthesis (Illumina; Table S4.1). A total of 13,682,363 reads for 3,360,832 unique *A. lyrata* small RNA, ranging in size from 15 to 30 nucleotides were generated and mapped to the *A. lyrata* genome, although most analyses used small RNA reads of 20 to 25 nucleotides. Like *A. thaliana*, *A. lyrata* small RNAs were mostly represented by 21 and 24-nucleotide RNA species, where the 21-nucleotide RNA overwhelmingly had a 5'U and the 24 nucleotide RNA were overrepresented with 5'A (Figure S4.2A). Small RNA-generating loci and reads mapped across each of the eight chromosomes, with enrichment around pericentromeric regions, similar to what was found for *A. thaliana* (Figures S4.1 and S4.2; Lu et al., 2005a; Lu et al., 2006; Rajagopalan et al., 2006; Kasschau et al., 2007). The density of 24-nucleotide small RNA loci was similar to the density pattern of transposable element loci and reciprocal to gene density, as in *A. thaliana* (Figure S4.1; Rajagopalan et al., 2006; Kasschau et al., 2007). By contrast, the density of 21-nucleotide generating loci was sparse, with discrete but abundant peaks, many of which corresponded to *MIRNA* and trans-acting siRNA (*TAS*) genes (Figure S4.1).

The numbers of small RNA loci that mapped to transposons, helitrons, long terminal repeat (LTR) and non-LTR retrotransposons, satellite/centromeric repeats, inverted repeats, and tandem repeats were uniformly higher in *A. lyrata* than in *A. thaliana*, although reads/million across each feature class did not show such a general bias (Figures S4.2B and S4.2C). The *A. lyrata* genome contains more repetitive elements than does the *A. thaliana* genome (392,271 repeats [62 Mb] versus 236,287 repeats [41 Mb], respectively; Table S4.2). After normalizing for feature class length, a similar density of small RNA reads/million/Mb was measured for most *A. thaliana* and *A. lyrata* repeat classes, although the read density from

LTR and non-LTR retrotransposons and tandem repeats was somewhat higher in *A. thaliana* (Figure S4.2D). Overall, the small RNA profiles are relatively similar between the two species.

Identification and conservation of *MIRNA* in *Arabidopsis* species

Previous studies identified at least 91 high-confidence *MIRNA* families in *A. thaliana*, as cataloged in miRBase release 14 (Griffiths-Jones et al., 2008). For reasons explained elsewhere (Axtell, 2008; Axtell and Bowman, 2008), *MIR401*, *MIR404-407*, *MIR413-420*, *MIR426*, *MIR782*, *MIR783*, *MIR854*, and *MIR855* were not included as bona fide *MIRNA* genes in this count. Among the 91 *MIRNA* families, homologs of nine were identified in the moss *Physcomitrella patens*, and up to 25 homologous families were identified in the angiosperms maize (*Zea mays*), rice (*Oryza sativa*), and poplar (*Populus spp*) (Axtell and Bowman, 2008; Zhang et al., 2009) (Figure 4.1). Two approaches were used to identify *MIRNA* genes in *A. lyrata*. First, an *MIRNA* orthology search was done between *A. thaliana* and *A. lyrata* genomes using MERCATOR and MAVID (Dewey, 2007). *A. thaliana MIRNA*s conserved in *A. lyrata* were defined as those at orthologous positions that were predicted to form self-complementary foldbacks with characteristics of canonical miRNA precursors (Meyers et al., 2008). Second, *MIRNA* were identified de novo using the *A. lyrata* small RNA sequencing data and computational filters using methods described previously (Fahlgren et al., 2007). For each previously unknown *A. lyrata MIRNA* gene identified by de novo search, the *A. thaliana* genome was inspected for a prospective orthologous locus. Orthologous *MIRNA* loci that contained mature miRNA sequences with four or more substitutions were defined as “diverged,” as this is at least twice the variation that has been used to identify conserved miRNA between distantly related species (Jones-Rhoades and Bartel, 2004). Additionally, each *A. lyrata* and *A. thaliana MIRNA* locus was used to search *Capsella rubella* genome sequences (represented by raw reads totaling ~307 Mb).

In total, 164 *A. lyrata MIRNA* loci were identified, representing 84 families (Figure 4.2). Read data, genomic data, and other information relevant to each MIRNA are given in Supplemental Data Set 1A and Supplemental Figure 3 online (http://www.plantcell.org). Of the 164 *MIRNA* loci, 101 (61.6%) yielded small RNA reads matching perfectly to the annotated mature miRNA and miRNA passenger strand (miRNA*), and 142 (86.6%) had at least one read matching either the annotated miRNA or miRNA* (see Supplemental Data Set 1A online; http://www.plantcell.org). Twenty-four *A. lyrata* families had at least two members, whereas 60 families were represented by only one member (Figure 4.2C). One hundred thirty-four *MIRNA* loci in *A. lyrata* had identifiable, conserved orthologous loci in *A. thaliana*, whereas 30 loci were either unique to *A. lyrata* or had diverged (Figures 4.2A and 4.2C). Seventeen previously unidentified *MIRNA* were discovered in *A. lyrata*, with two, *MIR774b* and *MIR3434*, having previously unrecognized orthologs in *A. thaliana*. Of 171 total *A. thaliana MIRNA* loci, 37 loci were either unique or diverged relative to *A. lyrata* (Figures 4.2A and 4.2C). These data indicate that a similar number, 18 and 22%, of *A. lyrata* and *A. thaliana MIRNA* loci, respectively, are either unique or substantially diverged. Given that *A. lyrata* is less well studied and has a larger genome, additional loci will likely be found in future studies.

The genomic sequences surrounding all *MIRNA* orthologs were aligned and compared using plots that highlighted conservation or divergence in both species (Figure 4.3). Most of the conserved orthologs, as well as most loci containing diverged mature miRNA loci, occurred in relatively colinear regions with relatively little rearrangement. This is illustrated by *MIR171a* and *MIR822* (conserved), as well as *MIR775*/*MIR3433* and *MIR402* (diverged) (Figures 4.3A and 4.3B). By contrast, comparison of the genomic regions surrounding *MIRNA* loci unique to either *A. lyrata* (*MIR3439*) or *A. thaliana* (*MIR843*) revealed, in nearly all cases, insertions or deletions at the orthologous region (Figure 4.3C). In 35 of 49 cases, these insertion/deletion events also included one or more adjacent, nonorthologous genes or transposons (see Supplemental Figure 4 online; http://www.plantcell.org). Local insertion-deletions, inversions, or duplications also accounted for each of the six *A. lyrata* loci (*MIR319d*, *MIR395g*, *MIR395h*, *MIR399g*, *MIR399h*, and *MIR399i*) containing species-specific paralogs for deeply conserved *MIRNA* families (Figure 4.3D).

Are unique *MIRNA* loci the result of species-specific additions or species-specific losses? To address this question, *A. lyrata* and *A. thaliana MIRNA* were compared with those in *C. rubella*, a species that belongs to a genus that is closely related to *Arabidopsis* within the Brassicaceae family (Koch et al., 2000). Unassembled reads from the *C. rubella* genome (National Center for Biotechnology Information Trace Archive; http://www.ncbi.nlm.nih.gov/Traces/home) were used to identify orthologs for known *Arabidopsis MIRNA* in *C. rubella*. Genomic reads (roughly representing 1X coverage of the genome) were assembled into small contigs using PCAP (Huang et al., 2003) and were aligned to the *A. thaliana* and *A. lyrata* genomes using AVID (Bray et al., 2003). In addition, *C. rubella* small RNA from seedlings and flowers were sequenced by high-throughput pyrosequencing (454 Life Sciences; 923,286 reads for 231,196 unique reads) and were mapped to the *C. rubella* contigs and singleton genomic reads (Table S4.1). In some cases, because of the low coverage of *C. rubella*, a *MIRNA* foldback was located at the edge of a contig, or spanned two contigs, but small RNA expression data still confirmed that the locus was active (see Supplemental Data Set 1B online; http://www.plantcell.org). It is recognized, however, that the *C. rubella* data do not represent full coverage of the genome. Based on the proportion of missing orthologs due to limited sequence coverage at loci corresponding to deeply conserved *MIRNA* families, it was estimated that up to 16% of the *C. rubella MIRNA* loci might be missing from the data set.

Despite the low coverage of the *C. rubella* genome, 112 *MIRNA* orthologs were identified (Figure 4.2A). Most (97) were conserved in both *Arabidopsis* species, and 14 were considered significantly diverged or orthologous to a *MIRNA* in only one *Arabidopsis* species (Figures 4.2A and 4.2B). Six loci were shared only between *C. rubella* and *A. lyrata*, and two were shared only between *C. rubella* and *A. thaliana*. The absence of these eight *MIRNA* loci in one of the *Arabidopsis* species is likely due to species-specific losses, and the greater number that has been lost in *A. thaliana* is consistent with the smaller genome size of this species. Seventeen *Arabidopsis MIRNA* loci (11 conserved and six diverged between the two species) were not detected in *C. rubella*. In addition, 10 of the loci unique to *A. thaliana* also lacked a *C. rubella* ortholog, as did nine of the *A. lyrata*–specific loci (Figure 4.2B). Due to the low coverage of the *C. rubella* genome, loci identified in one *Arabidopsis* species (22 loci; 16 in *A. thaliana*, six in *A. lyrata*) or both (22 loci; 21 conserved and one diverged) could not be confidently identified as present or absent in *C. rubella*. Given a 16% false negative rate estimate, seven of these loci are expected to be found in the *C. rubella genome*, and 37 are likely absent. Conservatively, at least 44 *MIRNA* families (58.3% of all *MIRNA* genes) are conserved between *Arabidopsis* species and *C. rubella*, nearly twice the number of families conserved between the three Brassicaceae species and the more distantly related dicot, *Populus trichocarpa* (Figures 4.1 and 4.2B).

Origins of *MIRNA* genes in *A. lyrata* and *A. thaliana*

The arms of foldbacks from several *A. thaliana MIRNA* genes have extended similarity (beyond just the miRNA and miRNA* sequence) with genes from target family members, which led to the hypothesis that new *MIRNA* families may arise by inverted duplication events involving sequences from what later become miRNA targets (Allen et al., 2004; Rajagopalan et al., 2006; Fahlgren et al., 2007; de Felippes et al., 2008). *MIRNA* loci with extended similarity to protein-coding gene sequences are generally considered to be young (Fahlgren et al., 2007; Axtell, 2008; Voinnet, 2009). All *MIRNA* gene foldbacks from *A. lyrata* and *A. thaliana* were evaluated for the presence of related sequences throughout the respective genomes. Among *MIRNA* from families conserved between *Arabidopsis* and *P. trichocarpa*, non-*MIRNA* sequences with significant similarity were detected for only one *MIRNA* (*MIR472*; Figures 4.4A and 4.4B). By contrast, of *MIRNA* conserved between *A. lyrata* and *A. thaliana*, but not *P. trichocarpa*, 36.4% exhibited significant similarity to at least one non-*MIRNA* locus

(Figures 4.4A and 4.4B). Similarly, 39.5% of *MIRNA* families unique to *A. thaliana* or *A. lyrata* displayed significant similarity to at least one non-*MIRNA* locus (Figures 4.4A and 4.4B; see Supplemental Data Set 1C online; http://www.plantcell.org).

The nature of each *MIRNA*-related locus in both genomes was investigated further to characterize the putative duplication events. All possible 12mers within 2-kb segments centered on each *MIRNA* gene and *MIRNA*-related locus were aligned to each other using BLAT (Kent, 2002), allowing for one mismatch. The relationship between the *MIRNA* and *MIRNA*-related loci was viewed by plotting connections between conserved 12mers (Figures 4.4C and 4.4D). Two general classes of duplications were detected. Approximately 79% of *MIRNA* families that had a *MIRNA*-related locus were predicted to have formed by an inverted duplication of the *MIRNA*-related locus (Figures 4.4C and 4.4E). The remaining 21% of *MIRNA* families appeared to be direct duplications of a *MIRNA*-related locus (Figures 4.4D and 4.4E). However, in the latter set, each *MIRNA*-related locus contained an ancestral inverted repeat that likely occurred before the *MIRNA*-forming duplication (Figure 4.4D). In many cases, the duplication extended well beyond the *MIRNA* foldback (Figures 4.4C and 4.4D). Regardless of duplication type, nearly one-half of the *MIRNA*-related loci were predicted to be targets of the corresponding miRNA (Figure 4.4E).

The nature of the *MIRNA*-related loci in *A. lyrata* and *A. thaliana* was examined as well. Over 82% of all *MIRNA*-related loci were protein-coding exon sequences (Figure 4.4F), and among these, ~57% were either predicted or validated to be targeted by the corresponding miRNA. This suggests that a substantial number of these young miRNAs have either lost targeting function or have evolved specificity to interact with a different target gene family, as proposed earlier (Fahlgren et al., 2007). One *MIRNA* appears to have originated from an intron sequence (*aly-MIR3444*), and nearly 12% had similarity to a nonannotated intergenic sequence. One *A. thaliana MIRNA* (*ath-MIR1888*) had similarity to numerous small inverted repeats, corresponding to a previously unknown family of miniature inverted-repeat transposable elements (MITEs; Figure 4.4F). These MITEs contain inverted repeats (~125 nucleotides) with flanking target site duplications.

Divergence of *A. lyrata* and *A. thaliana MIRNA* foldback sequences

Other than the mature miRNA, and to a lesser extent the miRNA*, *MIRNA* foldback sequences are not well conserved between distant lineages, making useful comparisons difficult (Jones-Rhoades et al., 2006). Taking advantage of the close relationship between the two *Arabidopsis* species, changes between orthologous *MIRNA* foldbacks from *A. thaliana* and *A. lyrata* were calculated. Normalized nucleotide divergence (substitutions per site) was measured between orthologous foldbacks that could be aligned confidently. Ninety-four orthologous pairs were from *MIRNA* families conserved to *P. trichocarpa*, whereas 32 pairs were from families not conserved to *P. trichocarpa*. Nucleotide divergence was measured independently for five foldback regions: miRNA sequence, miRNA* sequence, sequence between the stem base (or loop-distal) and the miRNA/miRNA* in the 5' arm (Region 1), the loop and loop-proximal sequences between the ends of the miRNA/miRNA*duplex (Region 2), and the sequence between the miRNA/miRNA* and the stem base in the 3' arm (Region 3; Figure 4.5A). As might be expected, nucleotide divergence was highest in loop-containing Region 2; the divergence in this region was not significantly different between the more conserved and the *Arabidopsis*-specific *MIRNA* (P = 0.493, permutation test; Figure 4.5B). Nucleotide divergence was intermediate at the base of the 3' arm in Region 3, with no significant difference in nucleotide divergence between the two sets (P = 0.862, permutation test; Figure 4.5B). However, in each of the other three regions, nucleotide divergence was significantly lower in the more conserved *MIRNA* gene set. This was particularly striking for the miRNA and miRNA* sequences ($P < 2 \times 10^{-16}$ and $P = 3.6 \times 10^{-5}$, respectively, permutation test; Figure 4.5B). These data strongly suggest that the young, *Arabidopsis*-specific *MIRNA* genes are under fewer evolutionary constraints than are the more deeply conserved *MIRNA* genes.

Variation at genomic loci flanking *MIRNA* genes in *A. lyrata* and *A. thaliana*

Are the recently evolved *MIRNA*, many of which originated by local duplication events, associated with more variable regions of the genome or with a higher density of transposable elements in *A. lyrata* and *A. thaliana*? The genomic environments surrounding *Arabidopsis MIRNA* genes were investigated in two ways. First, the lengths of the intergenic sequences flanking the 5' and 3' ends of each *MIRNA* locus, as well as those of all annotated genes, were measured and plotted in two dimensions (Haas et al., 2009). The density of transposons and repeat sequences, which are frequently associated with lower gene density, was also calculated for each intergenic space. *A. thaliana* genes were generally closer together than *A. lyrata* genes, with the median distance from another upstream or downstream gene being ~880 or ~1465 bp, respectively (Figure 4.6A). In *A. thaliana* and *A. lyrata*, the range of gene spacing that covered bins with eight or more genes was 67 to 9897 bp or 122 to 18,034 bp, respectively (Figure 4.6A). Most (88.5% of *A. thaliana* and 88.7% of *A. lyrata*) genes were spaced in this way. By contrast, the density of repeats and transposons was higher within longer intergenic spaces in both species (Figure 4.6B). In *A. thaliana*, repeat density was more concentrated in the largest intergenic regions (Figure 4.6B, left), whereas in *A. lyrata*, repeat density was relatively high in average to moderately large intergenic spaces (Figure 4.6B, right). This might indicate that repeats in *A. lyrata* are more evenly distributed, although this might be biased by the absence of ~17 Mb of centromeric sequence not included in the *A. lyrata* chromosome assemblies (http:// genome.jgi-psf.org/Araly1/Araly1.home.html). Regardless, in both species, bins with transposable elements and repeats occupying 30% or more of the intergenic space were highly enriched in regions with genes spaced further compared with the large majority of genes ($P < 2.2 \times 10^{-16}$, Fisher's exact test, bins in the boxed region versus bins above and to the right of the boxed region in Figure 4.6B). The distribution of intergenic distances adjacent to *MIRNA* genes was similar to the overall intergenic distances where 97% of *A. thaliana MIRNA* were within 67 to 9897 bp and 88.1% of *A. lyrata MIRNA* were within 122 to 18,034 bp from another upstream or downstream gene (Figure 4.6C). There was no difference between the lengths of intergenic spaces adjacent to *MIRNA*, whether or not they had orthologs or were unique to *A. thaliana* or *A. lyrata* (P = 0.09627 and P = 0.2179, respectively, Fisher's exact test; Figure 4.6C). *MIRNA* upstream and downstream regions were further examined by plotting the density of transposable elements and repeats in scrolling windows (window = 100 nucleotides, scroll = 20 nucleotides). Region metaplots were grouped by the depth of conservation of the *MIRNA* family to detect whether or not evolutionarily younger *MIRNA* were in regions with higher repeat density (Figure 4.6D). Repeat density metaplots for most groups did not appear different, as the numbers of repeats found near older *A. thaliana MIRNA* and all *A. lyrata MIRNA* were not significantly different (P > 0.05, Kruskal-Wallis rank sum test; Figure 4.6D). However, regions around *MIRNA* unique to *A. thaliana* had a significant enrichment of repeats relative to other *A. thaliana MIRNA* regions (P < 0.05, Kruskal-Wallis rank sum test; Figure 4.6D). Therefore, although the majority of *MIRNA* in both species were located in regions of normal gene density and relatively low repeat density (versus other genomic regions), very young *A. thaliana MIRNA* may be associated with more transposable elements and other repetitive sequences than older *MIRNA*.

In a second series of analyses, sequence variability adjacent to *MIRNA* genes in the *A. thaliana* and *A. lyrata* genomes was measured. For each *MIRNA* locus, the orthologous flanking regions (20,000 nucleotides upstream and downstream) inferred from the MERCATOR/MAVID alignment were extracted, and the numbers of unique, nonalignable positions due to insertions or deletions were quantified for each species. *MIRNA* genes were assigned to one of three conservation groups within each species: *MIRNA* family conserved to *P. trichocarpa* (Group 1); *MIRNA* conserved between *Arabidopsis* species, but not *P.*

trichocarpa (Group 2); and *MIRNA* unique to *A. lyrata* or *A. thaliana* (Group 3). In *A. thaliana*, there was no difference between Groups 1 and 2, although flanking regions of species-specific Group 3 *MIRNA* had significantly more unique nucleotides (P < 0.01, Kruskal-Wallis rank sum test; Figures 4.7A and 4.7B). *A. lyrata* Groups 1 and 2 *MIRNA*-adjacent regions contained more unique nucleotides than did *A. thaliana* Groups 1 and 2 *MIRNA*-adjacent regions (P < 0.01; Figures 4.7A and 4.7B). In *A. lyrata*, however, there was no significant difference in unique flanking nucleotides among the three conservation groups or between these groups and species-specific Group 3 *A. thaliana MIRNA* (Figures 4.7A and 4.7B). Collectively, these two analyses suggest that young *MIRNA* in *A. thaliana* may be associated with regions of relatively higher variability. *A. lyrata MIRNA* regions were similar in all conservation groups.

Evolution of miRNA targets

In *A. thaliana*, a set of 226 experimentally validated target transcripts, or transcripts for which high-confidence predictions have been made, was used to assess target site conservation in *A. lyrata* (see Supplemental Data Set 1D online; http://www.plantcell.org; most were also listed by The Arabidopsis Information Resource [TAIR] at ftp://ftp.arabidopsis.org/home/tair/Genes/SmallRNAs, on http://www.arabidopsis.org, January 12, 2010). Because more than one miRNA family targets some *A. thaliana* transcripts, the number of miRNA-target pairs (242) is greater than the number of target genes (see Supplemental Data Set 1D online; http://www.plantcell.org). Orthologs of *A. thaliana* miRNA targets were identified in *A. lyrata* using the MERCATOR/MAVID orthology map. The program TARGETFINDER (http://jcclab.science.oregonstate.edu/node/view/56334), which uses a score penalty system involving a set of consensus criteria, was used to assess miRNA target potential (Fahlgren et al., 2007). *A. lyrata* target orthologs with a score of 4 or less were considered conserved, but with some exceptions (see below).

Of the 242 *A. thaliana* miRNA-target pairs, 162 were conserved in *A. lyrata* (Figure 4.8A; see Supplemental Data Set 1D online; http://www.plantcell.org). Most (146) of the conserved target pairs had target prediction scores of 4 or less in both *A. thaliana* and *A. lyrata* (Figure 4.8A). An additional eight target pairs had target prediction scores in *A. lyrata* that were >4 but less than or equal to their validated *A. thaliana* orthologs and were therefore considered conserved as well (Figure 4.8A). In six cases, in which the *A. thaliana* target site was located in an untranslated region, a low-scoring *A. lyrata* target site was identified less than 270 nucleotides from the end of the *A. lyrata* ortholog, although in a nonannotated sequence. Lastly, *A. lyrata TAS3b* and *TAS3c* had target prediction scores of 8 and 4.5, respectively, but were considered conserved because of the deep conservation of the *TAS3* family and the high target prediction scores for these transcripts in *A. thaliana* (7 and 3.5, respectively; Figure 4.8A).

The remaining (80) *A. thaliana* miRNA-target pairs were not identified as conserved in *A. lyrata*. In half (40) of these, the target had no ortholog in *A. lyrata*, or the target site contained a disruptive insertion or deletion (Figure 4.8A). In some cases (10), the miRNA-target pair was not conserved because the miRNA itself was absent in *A. lyrata* (Figure 4.8A). The other miRNA-target pairs (29) represent potentially degraded or lost target sites, as the target ortholog in *A. lyrata* had a target prediction score >4 (and greater than the *A. thaliana* score; Figure 4.8A).

To determine if the target site variation between *A. thaliana* and *A. lyrata* correlated with conservation level of the corresponding *MIRNA*, miRNA-target pair categories were plotted individually for the three conservation groups defined in Figure 4.7. For the highly conserved *MIRNA* families (Group 1), 90% of the *A. thaliana* miRNA-target pairs were conserved in *A. lyrata* (Figure 4.8B). Most (64%) of the nonconserved targets in this category were mRNA from disease resistance genes (*CC-NBS-LRR*; miR472), which are among the most variable of gene classes in terms of major effect changes (Clark et al., 2007) (see Supplemental Data Set 1D online; http://www.plantcell.org). Most of the target differences were due to either the gain or loss of an ortholog or to a disruption of the target site sequence (Figure 4.8B). By contrast, for Group 2 miRNA conserved in both *Arabidopsis* species but not *P. trichocarpa*, only ~50% of the *A. thaliana* miRNA-target pairs were conserved in *A. lyrata* (Figure 4.8B). Approximately one-half of the differences were due to accumulation of point substitutions at a target site or in the mature miRNA, and one-half were from the absence of an ortholog or equivalent target site sequence (Figure 4.8B). Nonconserved target transcripts for Group 2 miRNA were primarily from the large *F-BOX* (42%; miR774 and miR859), *JACALIN-LIKE LECTIN* (19%; miR842 and miR846), and *PENTATRICOPEPTIDE REPEAT* (20%; miR161 and miR400) families (see Supplemental Data Set 1D online; http://www.plantcell.org). miRNA-target pairs involving Group 3 miRNA, which are specific to *A. thaliana*, by definition are all specific to *A. thaliana* (Figure 4.8B). These results suggest that whereas some young miRNA-target interactions may be conserved, many of these interactions are evolutionarily transient.

DISCUSSION

Origins of young *MIRNA*

The use of high-throughput sequencing has led to the discovery of large numbers of lineage-restricted *MIRNA* in diverse plant and algal species (Lu et al., 2006; Rajagopalan et al., 2006;

Axtell et al., 2007; Fahlgren et al., 2007; Molnar et al., 2007; Zhao et al., 2007; Heisel et al., 2008; Lu et al., 2008a; Morin et al., 2008; Moxon et al., 2008; Sunkar et al., 2008; Szittya et al., 2008; Zhu et al., 2008; Lelandais-Briere et al., 2009). In *A. thaliana*, only one-quarter of all of *MIRNA* families are conserved with *P. trichocarpa* or more distantly related species. The vast majority of *MIRNA* families show patterns consistent with more recent evolution. What are the rates of gain and loss of young *MIRNA* genes in the *Arabidopsis* lineage? It is difficult to measure birth and death rates directly because the presence and absence of a gene in two extant species could be interpreted as a gain in one or a loss in the other. In some cases the presence of the gene in an outgroup species parsimoniously indicates that the gene was lost in one lineage, as was the case with several *MIRNA* found in *C. rubella* and either *A. thaliana* (*MIR830* and *MIR865*) or *A. lyrata* (*MIR395g* and *h*, *MIR399g-l*, and *MIR3435*). Instead of estimating the birth and death rates directly, the net flux, or composite of births and deaths, can be estimated from the extant *MIRNA* identified in *A. thaliana* or *A. lyrata* but missing from *C. rubella*. Due to the incomplete sequence available for the *C. rubella* genome, estimates of the rate of *MIRNA* flux in the *Arabidopsis* lineage based on *MIRNA* families confidently identified as absent in *C. rubella* will be conservative. A liberal estimate of the rate of *MIRNA* flux can be estimated from the missing *C. rubella* data, adjusting for a 16% false negative rate. Conservatively, 24 and 25 *MIRNA* families were identified in *A. thaliana* and *A. lyrata*, respectively, but not *C. rubella*. Liberally, 46 and 40 *MIRNA* families were identified in *A. thaliana* and *A. lyrata*, respectively, and may not be present in *C. rubella*. Therefore, assuming ~20 million years of divergence between *C. rubella* and *Arabidopsis* species (Koch et al., 2000; Wright et al., 2002; Ossowski et al., 2010), the rate of flux of *Arabidopsis MIRNA* families is conservatively 1.2 to 1.3, or liberally 2.0 to 2.3 genes per million years. An additional 31 *A. lyrata MIRNA* families that did not overlap the *MIRNA* identified here were identified in an independent study (Ma et al., 2010). If these loci are included in the estimate, then the rate of flux of *Arabidopsis MIRNA* families could be as high as 3.3 genes per million years. A recent study of drosophilid *MIRNA* reported a rate of flux of 0.82 to 1.6 genes per million years (Berezikov et al., 2010), which overlaps with the conservative *Arabidopsis* species estimates.

How are new *MIRNA* genes forming? Based on sequence similarity searches against the *A. thaliana* and *A. lyrata* genomes, a large proportion of *MIRNA* originated from intragenomic duplications of protein-coding genes. By expanding the alignment to regions flanking the *MIRNA* foldback and foldback-similar region, we detected extended locus similarity in many cases (up to 2 kb), suggesting that *MIRNA* loci can form from larger duplication events. *MIRNA*-related loci were found for more than one-third of *MIRNA* genes conserved between, or unique to, *A. thaliana* and *A. lyrata*. What processes formed the large number of loci that show no evidence of duplication-driven origin? At least some of these *MIRNA* may have lost extended similarity to their originating locus due to sequence divergence or loss of the originating locus. Loss of similarity becomes more likely as the *MIRNA* locus ages. Another possibility is that some *MIRNA* are formed from random self-complementary regions or other types of features that have a self-complementary nature. The identification of intergenic- (*MIR843*, *MIR849*, *MIR850*, and *MIR863*) and MITE-derived (*MIR1888*) *MIRNA* supports this idea. Loci like *MIR1888* were identified because of the presence of related MITEs in the genome, but the formation of *MIRNA* from random self-complementary regions would not necessarily require duplication events. Rather, these regions could acquire *MIRNA* features through random mutation of the original locus (Chen and Rajewsky, 2007; de Felippes et al., 2008).

If at least some *MIRNA* loci are formed through duplication and rearrangement events, are young *MIRNA* associated with more variable regions of the genome? Based on examination of the most recently evolved group of *A. thaliana MIRNA*, there was a strong regional association with unique, or unaligned, sequences, compared with *MIRNA* from families conserved in *A. lyrata* or conserved more deeply. Additionally, young *A. thaliana MIRNA* may be associated with more transposons. However, analysis of *A. lyrata MIRNA* was less clear because all conservation groups were associated with flanking regions containing similar amounts of unique sequence. Comparative analysis of *A. thaliana*, *A. lyrata*, and other close relatives suggests that *A. thaliana* has evolved a reduced genome size through both large and small deletion events (Oyama et al., 2008). These deletions, which appear as unaligned nucleotides in *A. lyrata*, may mask patterns of variability found near young *A. lyrata MIRNA*. Future comparisons between *A. lyrata* and additional *Arabidopsis* species that share the ancestral genome architecture will be helpful in elucidating this possibility.

Diversification of *MIRNA*

Is the comparative analysis of *A. thaliana* and *A. lyrata* informative about the functionality of recently evolved *MIRNA*? As a group, younger miRNA are significantly more divergent than deeply conserved miRNA. This suggests that purifying selection is acting on the most deeply conserved *MIRNA*, as noted by others (Ehrenreich and Purugganan, 2008; Warthmann et al., 2008). Understanding why young miRNA are more diverse is less clear. Sequence diversification could be the result of neutral mutational drift or non-neutral evolution in one or both species. Neutral evolution would imply low or no functionality, while non-neutral evolution could purge *MIRNA* to avoid deleterious miRNA-target interactions. Although nucleotide divergence between young *MIRNA* was significantly higher than between deeply conserved *MIRNA*, divergence in the miRNA region was significantly lower than in the loop-proximal region (Region 2), suggesting that at least some young miRNA may be evolutionarily constrained. However, the generally lower level of expression of young miRNA supports the idea that drift may be the primary evolutionary force acting on these loci (Axtell, 2008), and in our *A. lyrata* data sets, the median expression level of miRNA from families conserved with *P. trichocarpa* was 5 times higher than for younger miRNA (see Supplemental Data Set 1A online; http://www.plantcell.org). In addition to young *MIRNA* being more diverged, miRNA-target interactions involving younger miRNA were also more divergent. About half of the miRNA-target pairs from young *MIRNA* were conserved between *A. thaliana* and *A. lyrata* versus 90% of pairs involving deeply conserved miRNA. For young *MIRNA* families, divergence of miRNA and target site sequences reflects the fluidity of targeting between *A. thaliana* and *A. lyrata*. Together, these data indicate that most young *MIRNA* may be evolving neutrally, with little or no functional consequence.

MATERIALS AND METHODS

Small RNA data sets and processing

Small RNA samples were extracted from wild type *Arabidopsis lyrata* MN47, *Arabidopsis thaliana* Columbia-0, and *Capsella rubella* MTE as by Fahlgren et al., (2009). Small RNA libraries from *A. lyrata* flower (stage 1-12) and 14-d-old seedlings (two libraries each) and *C. rubella* flower (stage 1-12), 14-d-old seedlings, 5-d-old seedlings treated for 6 h with Murashige and Skoog broth plus 150 mM NaCl and 5-d-old seedlings mock treated with Murashige and Skoog broth were constructed and sequenced by pyrosequencing (454 Life Sciences) in a multiplexed format as by Kasschau et al., (2007). Samples were barcoded with a unique 5' adaptor: *A. lyrata* flower sample 1 (5'-ATCGTAG**CGCA**CUGAUA-3'), flower sample 2 (5'-ATCGTAG**CGAC**CUGAUA-3'), seedling sample 1 (5'-ATCGTAG**CGUG**CUGAUA-3'), and seedling sample 2 (5'-ATCGTAG**GCGU**CUGAUA-3'); *C. rubella* flower sample (5'-ATCGTAG**CGCA**CUGAUA-3'), 14 d seedling (5'-ATCGTAG**CGUG**CUGAUA-3'), 6 h NaCl seedling (5'-ATCGTAG**CGAC**CUGAUA-3'), and 6 h mock seedling (5'-ATCGTAG**GCGU**CUGAUA-3') where the unique barcode is in bold (Kasschau et al., 2007). Additionally, small RNA libraries from *A. lyrata* rosette leaves (one library) and *A. thaliana* total aerial tissue 21 d (one library) were constructed and sequenced using sequencing-by-synthesis (Illumina) as described (Fahlgren et al., 2009). Libraries for *A. lyrata* flower (stage 1-12; two libraries) were constructed as by Mosher et al., (2009), except small RNA were isolated by PAGE and RNA amplicons were reverse transcribed using the

Fermentas Revertaid kit (Fermentas Life Sciences) and amplified by PCR using the Phusion DNA polymerase (Finnzymes) and then sequenced using sequencing-by-synthesis (Illumina). After sequencing, all data were processed and mapped to their respective genome (*A. lyrata* [v1.0; http://genome.jgi-psf.org/Araly1/Araly1.home.html], *A. thaliana* [TAIR8], or *C. rubella* raw reads and assembled contigs) using the CASHX pipeline as described (Fahlgren et al., 2009). *A. lyrata* 454 libraries, the *A. lyrata* Illumina leaf library, and the *A. thaliana* Illumina library were used for comparisons in Figure S4.2. *A. lyrata* 454 libraries and the Illumina leaf library were used to identify *A. lyrata MIRNA* loci by de novo search. All *A. lyrata* libraries were used to evaluate de novo *MIRNA* predictions and read counts listed in Supplemental Data Set 1A online (http://www.plantcell.org).

Repeat masking

Repeat elements in the *A. lyrata* genome (v1.0) were identified using RepeatMasker (v open-3.2.5; http://www.repeatmasker.org). Repeat libraries for *Arabidopsis* were used to identify *A. lyrata* repeats (species *Arabidopsis*) with default settings. Tandem repeats were identified using Tandem Repeat Finder (v 4.0 linux-64 bit; Benson, 1999) with match weight = 2, mismatch penalty = 3, indel penalty = 5, match probability = 80, indel probability = 10, minimum alignment score = 40, and maximum period size = 1000. Inverted repeats were identified using Inverted Repeat Finder (v 3.05 linux-64 bit; Warburton et al., 2004) with match weight = 2, mismatch penalty = 3, indel penalty = 5, match probability = 80, indel probability = 10, minimum alignment score = 40, maximum stem length = 100,000, and maximum loop length = 500,000.

Whole-genome alignment

Orthologous regions of the five and eight nuclear chromosomes of *A. thaliana* and *A. lyrata*, respectively, were identified and aligned using MERCATOR and MAVID (Dewey, 2007). Briefly, interspersed and low complexity repeats were used to make hard- and soft-masked versions of each genome, respectively. The masked and unmasked versions of the genomes were converted to SDB format for use in the MERCATOR pipeline. Coding sequence annotation for *A. thaliana* (TAIR8, http://www.arabidopsis.org; Swarbreck et al., 2008) and *A. lyrata* (filtered gene models; Joint Genome Initiative, http://genome.jgi-psf.org/Araly1/Araly1.home.html) were used to create a preliminary orthology map of the two genomes. The orthology map was refined by combining mapping intervals between consecutive regions and by determining breakpoints between consecutive regions containing rearrangement. Finally, the orthology map was used to align the genomes using MAVID (Bray and Pachter, 2004) with the assumed phylogenetic relationship (A.thaliana: 0.0321, A.lyrata:0.0257) from Hoffmann, (2005).

Identification of *A. lyrata MIRNA*

Orthologs and paralogs of *A. thaliana MIRNA* were identified as stated above. Previously unknown *A. lyrata MIRNA* were identified by a de novo computational pipeline similar to that used in for *A. thaliana* (Fahlgren et al., 2007). Small RNA (20 to 22 nucleotides) that were represented by two or more reads among all libraries, that matched the genome 10 or fewer times and that did not overlap structural RNA genes or repetitive loci, were used initially to seed the pipeline. An initial foldback scan was done with small RNA-flanking sequence using Inverted Repeat Finder (Warburton et al., 2004) and RNAFOLD (Vienna RNA package v1.81; Hofacker, 2003). Overlapping foldbacks for candidates were consolidated. Foldback structures that were associated with small RNA in which at least 95% were from the foldback polarity were retained. Next, a minimal foldback was predicted using a Perl script to iteratively run RNAFOLD, with foldback sequence trimming on each cycle until only a single stem-loop structure remained with the predicted miRNA and miRNA* sequences in canonical precursor context (Meyers et al., 2008). The small RNA database was screened for predicted miRNA* sequences with 2-nucleotide 3' overhangs relative to the predicted miRNA. Features of each *A. lyrata MIRNA* locus are provided in Supplemental Data Set 1B online (http://www.plantcell.org).

MIRNA-related locus analysis

All *MIRNA* foldbacks from *A. thaliana* and *A. lyrata* were aligned against the *A. thaliana* and *A. lyrata* genome, respectively, with FASTA (v34) (Pearson, 1990). *MIRNA* sequences were masked from the genomes to prevent self-matches. The –log of the top four expect (E) values was plotted for each *MIRNA* (Figure 4.4A). The E-values were converted to P values using the relationship $P = 1\text{-}e^{-E}$, and an FDR cutoff point was determined using the R (v2.9.2; R Development Core Team, 2009) Q-VALUE package (v1.0; Storey, 2002). For each significant (FDR ≤ 0.05) alignment pair, a flanking region (1 kb upstream and downstream) around each locus was computationally shredded into all possible overlapping 12-nucleotide fragments. Fragments from the *MIRNA* locus were aligned to fragments from the *MIRNA*-related locus with BLAT (Kent, 2002), allowing for one mismatch and no gaps (-t=dna –q=dna –tileSize=12 –stepSize=1 –oneOff=1 –minMatch=1 –minScore=9 –maxGap=0).

Nucleotide divergence and statistical analyses

Orthologous pairs of *A. thaliana* and *A. lyrata MIRNA* were first sorted to a conservation group based on whether or not the family was conserved in *P. trichocarpa*. Nucleotide divergence at orthologous *MIRNA* pairs was done after an initial alignment with ClustalW (v1.83; Thompson et al., 1994) with the following settings: -type=DNA -dnamatrix=IUB -gapopen= 10 -gapext=0.2 -gapdist=8 -transweight=0.5 -endgaps -ktuple=1 -output= fasta. *MIRNA* foldback alignments were parsed into five regions: miRNA sequence; miRNA* sequence; sequence between the stem base (or loop-distal) and the miRNA/miRNA* in the 5' arm (Region 1); the loop and loop-proximal sequences between the ends of the miRNA/miRNA*duplex (Region 2); and the sequence between the miRNA/miRNA* and the stem base in the 3' arm (Region 3; Figure 4.5). Polymorphisms were counted within each region of orthologous pairs (SNP count + indel count, where a contiguous indel counted as one polymorphism). Substitutions per site were calculated by dividing the total polymorphisms by the length (nucleotides) of the region. A permutation test with one million simulations was done to test for significant differences in substitutions/site within regions and between conservation groups (twot.permutation function in the DAAG package (Maindonald and Braun, 2007) in R v2.9.2; R Development Core Team, 2009).

To analyze unique positions flanking *A. thaliana* and *A. lyrata MIRNA* (Figures 4.7A and 4.7B), 20,000 nucleotides upstream and downstream of each *MIRNA* in both species was extracted from the MERCATOR/MAVID alignment. Unique positions were defined as unaligned (gapped alignment) nucleotides. Boxplots of unique nucleotides were generated using the boxplot function in the R graphics package (R Development Core Team, 2009). A nonparametric analysis of variance test (Kruskal-Wallis rank sum test) was done to assess differences between groups using the kruskal.test function (R stats package; R Development Core Team, 2009). Corrected, significant pairwise differences were determined using a multiple comparison test after Kruskal-Wallis with the kruskalmc function (pgirmess v1.3.8 R package).

Accession numbers

Small RNA data sets used here were deposited in the National Center for Biotechnology Information's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through the series accession GSE20662 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20662). All *MIRNA* loci were deposited at miRBase (http://www.mirbase.org; Griffiths-Jones et al., 2008), including the previously undiscovered accessions *MIR3433* to *MIR3449*.

ACKNOWLEDGMENTS

We thank the U.S. Department of Energy Joint Genome Institute for producing the *A. lyrata* and *C. rubella* genome sequence under the Community Sequencing Program. We thank Christa Lanz, Korbinian Schneeberger, and Stephan Ossowski for help with Illumina sequencing. We thank members of the Carrington and Weigel labs for productive discussions. We also thank Goretti Nguyen for assistance with small RNA library preparation. N.F. was supported in part by a P.F. and Nellie Buck Yerex Fellowship. L.M.S. was supported by European Community FP7 Marie Curie Fellowship (PIEF-GA-2008-221553). Grant support for this work in the Carrington lab came from the National Science Foundation (MCB-0618433) and the National Institutes of Health (AI43288), and in the Weigel lab from European Community FP6 IP SIROCCO (Contract LSHG-CT-2006-037900) and the Max Planck Society. This work was greatly facilitated by a Humboldt research award from the Alexander von Humboldt Foundation to J.C.C. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy (Contract DE-AC02-05CH11231).

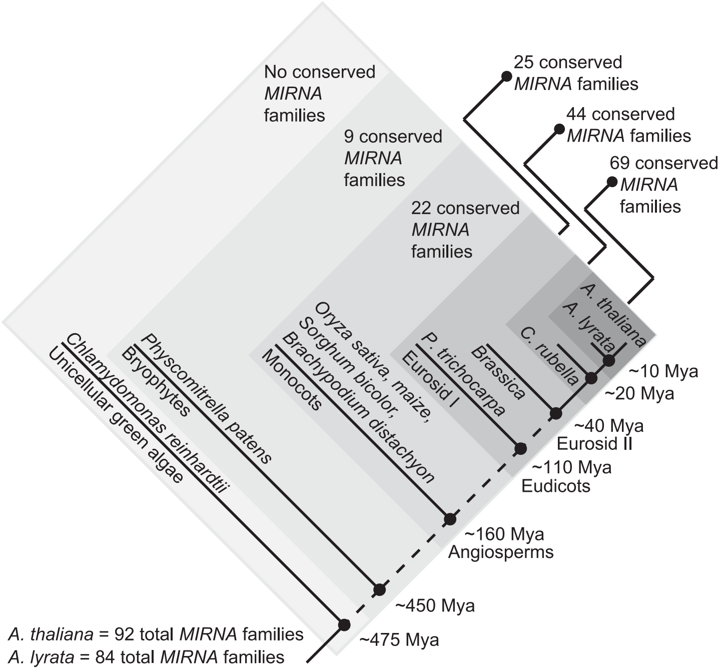
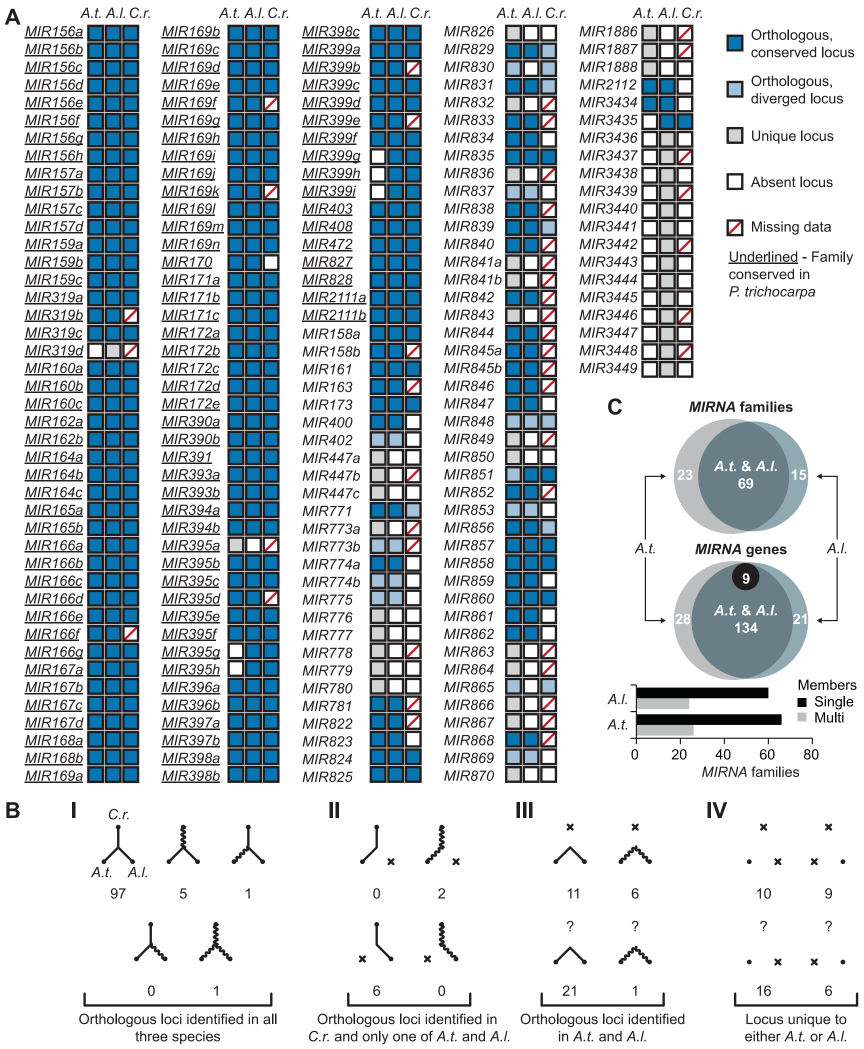

No conserved *MIRNA* families
9 conserved *MIRNA* families
22 conserved *MIRNA* families
25 conserved *MIRNA* families
44 conserved *MIRNA* families
69 conserved *MIRNA* families
A. thaliana
A. lyrata
C. rubella
Brassica
P. trichocarpa
Eurosid I
Oryza sativa, maize, Sorghum bicolor, Brachypodium distachyon
Monocots
Physcomitrella patens
Bryophytes
Chlamydomonas reinhardtii
Unicellular green algae
~10 Mya
~20 Mya
~40 Mya
Eurosid II
~110 Mya
Eudicots
~160 Mya
Angiosperms
~450 Mya
~475 Mya
A. thaliana = 92 total *MIRNA* families
A. lyrata = 84 total *MIRNA* families

Figure 4.1. Conservation of *Arabidopsis MIRNA* families in plants.
The cladogram (not drawn to scale) represents the plant tree of life with major phylogenetic groups noted. Each box is labeled with the inferred number of *Arabidopsis MIRNA* families conserved between all available taxa within the box, based on the representative species listed. The numbers of *MIRNA* families in *A. thaliana* and *A. lyrata* are listed at the bottom left.



	A.t.	A.l.	C.r.
MIR156a	🟦	🟦	🟦
MIR156b	🟦	🟦	🟦
MIR156c	🟦	🟦	🟦
MIR156d	🟦	🟦	🟦
MIR156e	🟦	🟦	🟦
MIR156f	🟦	🟦	🟦
MIR156g	🟦	🟦	🟦
MIR156h	🟦	🟦	🟦
MIR157a	🟦	🟦	🟦
MIR157b	🟦	🟦	🟦
MIR157c	🟦	🟦	🟦
MIR157d	🟦	🟦	🟦
MIR159a	🟦	🟦	🟦
MIR159b	🟦	🟦	🟦
MIR159c	🟦	🟦	🟦
MIR319a	🟦	🟦	🟦
MIR319b	🟦	🟦	⧄
MIR319c	🟦	🟦	🟦
MIR319d	□	▒	⧄
MIR160a	🟦	🟦	🟦
MIR160b	🟦	🟦	🟦
MIR160c	🟦	🟦	🟦
MIR162a	🟦	🟦	🟦
MIR162b	🟦	🟦	🟦
MIR164a	🟦	🟦	🟦
MIR164b	🟦	🟦	🟦
MIR164c	🟦	🟦	🟦
MIR165a	🟦	🟦	🟦
MIR165b	🟦	🟦	🟦
MIR166a	🟦	🟦	🟦
MIR166b	🟦	🟦	🟦
MIR166c	🟦	🟦	🟦
MIR166d	🟦	🟦	🟦
MIR166e	🟦	🟦	🟦
MIR166f	🟦	🟦	⧄
MIR166g	🟦	🟦	🟦
MIR167a	🟦	🟦	🟦
MIR167b	🟦	🟦	🟦
MIR167c	🟦	🟦	🟦
MIR167d	🟦	🟦	🟦
MIR168a	🟦	🟦	🟦
MIR168b	🟦	🟦	🟦
MIR169a	🟦	🟦	🟦
MIR169b	🟦	🟦	🟦
MIR169c	🟦	🟦	🟦
MIR169d	🟦	🟦	🟦
MIR169e	🟦	🟦	🟦
MIR169f	🟦	🟦	⧄
MIR169g	🟦	🟦	🟦
MIR169h	🟦	🟦	🟦
MIR169i	🟦	🟦	🟦
MIR169j	🟦	🟦	🟦
MIR169k	🟦	🟦	⧄
MIR169l	🟦	🟦	🟦
MIR169m	🟦	🟦	🟦
MIR169n	🟦	🟦	🟦
MIR170	🟦	🟦	□
MIR171a	🟦	🟦	🟦
MIR171b	🟦	🟦	🟦
MIR171c	🟦	🟦	🟦
MIR172a	🟦	🟦	🟦
MIR172b	🟦	🟦	🟦
MIR172c	🟦	🟦	🟦
MIR172d	🟦	🟦	🟦
MIR172e	🟦	🟦	🟦
MIR390a	🟦	🟦	🟦
MIR390b	🟦	🟦	🟦
MIR391	🟦	🟦	🟦
MIR393a	🟦	🟦	🟦
MIR393b	🟦	🟦	🟦
MIR394a	🟦	🟦	🟦
MIR394b	🟦	🟦	🟦
MIR395a	▒	□	⧄
MIR395b	🟦	🟦	🟦
MIR395c	🟦	🟦	🟦
MIR395d	🟦	🟦	⧄
MIR395e	🟦	🟦	🟦
MIR395f	🟦	🟦	🟦
MIR395g	□	🟦	🟦
MIR395h	□	🟦	🟦
MIR396a	🟦	🟦	🟦
MIR396b	🟦	🟦	🟦
MIR397a	🟦	🟦	🟦
MIR397b	🟦	🟦	🟦
MIR398a	🟦	🟦	🟦
MIR398b	🟦	🟦	🟦
MIR398c	🟦	🟦	🟦
MIR399a	🟦	🟦	🟦
MIR399b	🟦	🟦	⧄
MIR399c	🟦	🟦	🟦
MIR399d	🟦	🟦	🟦
MIR399e	🟦	🟦	⧄
MIR399f	🟦	🟦	🟦
MIR399g	□	🟦	🟦
MIR399h	□	🟦	🟦
MIR399i	□	🟦	🟦
MIR403	🟦	🟦	🟦
MIR408	🟦	🟦	🟦
MIR472	🟦	🟦	🟦
MIR827	🟦	🟦	🟦
MIR828	🟦	🟦	🟦
MIR2111a	🟦	🟦	🟦
MIR2111b	🟦	🟦	🟦
MIR158a	🟦	🟦	🟦
MIR158b	🟦	🟦	⧄
MIR161	🟦	🟦	🟦
MIR163	🟦	🟦	⧄
MIR173	🟦	🟦	🟦
MIR400	🟦	🟦	□
MIR402	▨	▨	□
MIR447a	▒	□	□
MIR447b	▒	□	⧄
MIR447c	▒	□	□
MIR771	🟦	🟦	▨
MIR773a	▒	□	⧄
MIR773b	▨	▨	⧄
MIR774a	🟦	🟦	□
MIR774b	▨	▨	□
MIR775	▨	▨	□
MIR776	▒	□	□
MIR777	▒	□	□
MIR778	▒	□	⧄
MIR779	▒	□	□
MIR780	▒	□	□
MIR781	🟦	🟦	⧄
MIR822	🟦	🟦	⧄
MIR823	🟦	🟦	□
MIR824	🟦	🟦	🟦
MIR825	🟦	🟦	🟦
MIR826	▒	□	□
MIR829	🟦	🟦	▨
MIR830	▨	□	▨
MIR831	🟦	🟦	▨
MIR832	▒	□	⧄
MIR833	🟦	🟦	⧄
MIR834	🟦	🟦	□
MIR835	🟦	🟦	🟦
MIR836	▒	□	⧄
MIR837	▨	▨	□
MIR838	🟦	🟦	⧄
MIR839	🟦	🟦	▨
MIR840	🟦	🟦	⧄
MIR841a	▒	□	⧄
MIR841b	▒	□	⧄
MIR842	🟦	🟦	⧄
MIR843	▒	□	⧄
MIR844	🟦	🟦	⧄
MIR845a	🟦	🟦	⧄
MIR845b	🟦	🟦	⧄
MIR846	🟦	🟦	⧄
MIR847	🟦	🟦	□
MIR848	▨	▨	▨
MIR849	▒	□	⧄
MIR850	▒	□	□
MIR851	▨	🟦	🟦
MIR852	🟦	🟦	⧄
MIR853	▨	▨	□
MIR856	🟦	🟦	▨
MIR857	🟦	🟦	🟦
MIR858	🟦	🟦	🟦
MIR859	🟦	🟦	□
MIR860	🟦	🟦	🟦
MIR861	🟦	🟦	□
MIR862	🟦	🟦	□
MIR863	▒	□	⧄
MIR864	▒	□	⧄
MIR865	▨	□	▨
MIR866	▒	□	⧄
MIR867	▒	□	⧄
MIR868	🟦	🟦	⧄
MIR869	▨	▨	□
MIR870	▒	□	□
MIR1886	▒	□	⧄
MIR1887	▒	□	⧄
MIR1888	▒	□	□
MIR2112	🟦	🟦	□
MIR3434	🟦	🟦	□
MIR3435	□	🟦	🟦
MIR3436	□	▒	□
MIR3437	□	▒	⧄
MIR3438	□	▒	□
MIR3439	□	▒	⧄
MIR3440	□	▒	□
MIR3441	□	▒	□
MIR3442	□	▒	⧄
MIR3443	□	▒	□
MIR3444	□	▒	□
MIR3445	□	▒	□
MIR3446	□	▒	⧄
MIR3447	□	▒	□
MIR3448	□	▒	⧄
MIR3449	□	▒	□

🟦 Orthologous, conserved locus

▨ Orthologous, diverged locus

▒ Unique locus

□ Absent locus

⧄ Missing data

Underlined - Family conserved in *P. trichocarpa*

A
B
I
C.r.
A.t.
A.l.
97
5
1
0
1
Orthologous loci identified in all three species
II
0
2
6
0
Orthologous loci identified in C.r. and only one of A.t. and A.l.
III
11
6
?
?
21
1
Orthologous loci identified in A.t. and A.l.
IV
10
9
?
?
16
6
Locus unique to either A.t. or A.l.
C
MIRNA families
23
A.t. & A.l.
69
15
A.t.
A.l.
MIRNA genes
9
28
A.t. & A.l.
134
21
A.l.
A.t.
Members
Single
Multi
0
20
40
60
80
MIRNA families

Figure 4.2. Orthology of *MIRNA* genes in *A. thaliana*, *A. lyrata*, and *C. rubella*. **(A)** Ortholog conservation matrix between *A. thaliana* (*A.t.*), *A. lyrata* (*A.l.*), and *C. rubella* (*C.r.*). *MIRNA* genes with orthologs are shown by a dark-blue box or as a light-blue box in the cases where the mature miRNA sequence has diverged in sequence. Lack of an ortholog is indicated as a white box. *MIRNA* genes found in only one species are shown with a gray box. *C. rubella MIRNA* that could not be confidently identified as present or absent are shown with a red slash. Genes with underlined names are from families conserved in *P. trichocarpa*. Note that the *A. lyrata* ortholog of *ath-MIR775* is named *aly-MIR3433*. **(B)** Conservation scenarios. Species vertices are labeled in the first diagram as in **(A)**. Orthologous *MIRNA* genes are depicted as dots connected by solid lines. Dots connected by undulating lines indicate a mature miRNA that has diverged in sequence. Absent or missing orthologs are depicted as "X." Scenarios under a question mark indicate that insufficient *C. rubella* data are available. The number of *MIRNA* observed for each scenario is listed below each diagram. **(C)** Summary of *MIRNA* families and genes in *A. lyrata* and *A. thaliana*. The number of *MIRNA* families or genes conserved between (overlap region), or unique to, *A. thaliana* (gray region) and *A. lyrata* (blue region) are shown (Venn diagrams, top). The black region is the number of diverged orthologs (Venn diagrams, bottom). The numbers of multi- and single-gene *MIRNA* families are shown for both species (bar chart).

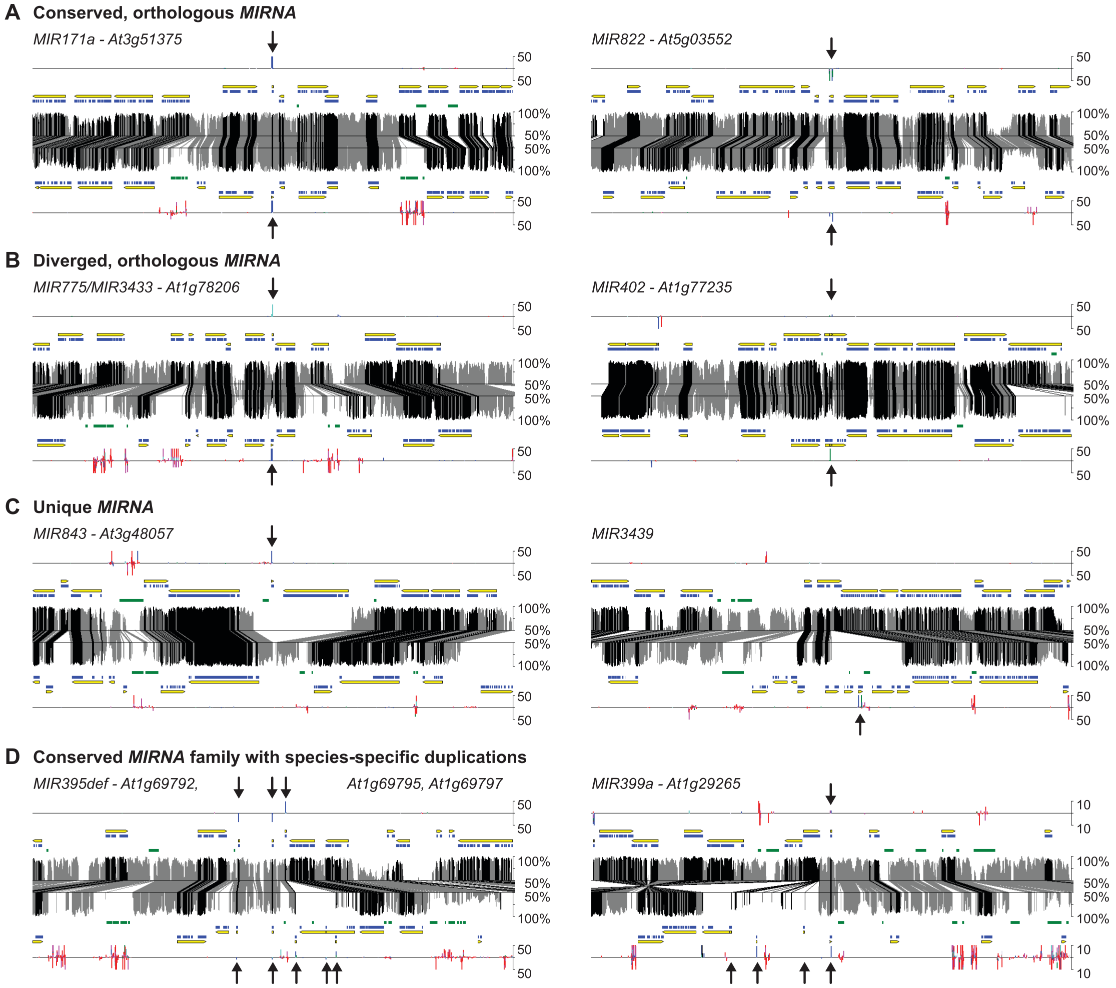


Figure 4.3. Locus conservation around several classes of *MIRNA* genes.
All panels show 40-kb regions flanking *MIRNA* genes. Each illustration shows the *A. thaliana* genome in the top half and the *A. lyrata* genome in the bottom half. Small RNA abundance at each nucleotide position for each strand (Watson strand reads are plotted up and Crick strand reads are plotted down) is shown in the top and bottom histograms. Small RNA reads are color coded by length (turquoise = 20 nucleotides, blue = 21 nucleotides, green = 22 nucleotides, fuchsia = 23 nucleotides, and red = 24 nucleotides). Genes and exons are drawn as yellow chevrons and blue boxes, respectively. Transposable elements are drawn as green boxes. Nucleotide conservation is illustrated for both genomes with orthologous positions linked. Conservation (percentage) was calculated using the scrolling window method (100-nucleotide windows, 1-nucleotide scroll) and plotted as a histogram with black bars representing positions that overlapped exons in *A. thaliana* and gray bars representing all other sites. Arrows in each panel mark the positions of the *MIRNA* gene indicated in the heading. **(A)** Conserved, orthologous *MIRNA*. The deeply conserved *MIR171a* (left) and *Arabidopsis*-specific *MIR822* (right) are shown. **(B)** Orthologous *MIRNA* genes with diverged mature miRNA sequences. **(C)** Unique *MIRNA*. *MIR843* and *MIR3439* are specific to *A. thaliana* and *A. lyrata*, respectively. **(D)** Conserved families with species-specific, duplicated (paralogous) *MIRNA* genes. *MIR395e* and *MIR395f* are duplicated in *A. lyrata* (left). *MIR399a* has three additional copies in *A. lyrata* (right).

Figure 4.4. Identification of intragenomic loci with extended similarity to *MIRNA* genes. **(A)** Detection of *MIRNA*-related loci in the *A. thaliana* and *A. lyrata* genomes. The expected values for the top four FASTA alignments between each *MIRNA* foldback and the respective genome sequence are plotted. *A. thaliana MIRNA*s aligned to the *A. thaliana* genome are plotted as red dots. *A. lyrata MIRNA*s aligned to the *A. lyrata* genome are plotted as black dots. *MIRNA* are grouped on the x axis based on conservation. The horizontal gray line marks the false discovery rate (FDR) = 0.05 boundary, where points above the line have a FDR < 0.05. **(B)** Proportion of *A. thaliana* and *A. lyrata MIRNA* families conserved with both *Arabidopsis* species and *P. trichocarpa*, conserved between *A. thaliana* and *A. lyrata* but not with *P. trichocarpa*, or *MIRNA*s unique to either *A. thaliana* or *A. lyrata* (inner pie chart). The percentage of *MIRNA* families with evidence of foldback origination from another locus (see **[A]**) in each conservation group is shown in the outer ring. **(C)** *MIRNA* formation by inverted duplication. A generic model (top) and three example *MIRNA* loci that were formed by inverted duplication events. **(D)** *MIRNA* formation by direct duplication of a locus containing a previous intralocus inverted duplication. A generic model (top) and three examples from *A. lyrata* are shown. In **(C)** and **(D)**, the *MIRNA*-related genomic locus is on top and the *MIRNA* genomic locus is on the bottom of each diagram. Direct and inverted duplications are shown by gray and blue shadings, respectively, connecting the *MIRNA*-related locus to the *MIRNA* locus. Annotated features are shown as highlighted boxes in each region (genes = yellow, exons = blue, transposable elements = green, and *MIRNA* = red boxes). Arrows indicate the duplicated segments in each locus. In **(D)**, sequences produced by inverted intralocus duplication in the *MIRNA*-related loci are connected by black boxes. **(E)** Proportion of *A. thaliana* and *A. lyrata MIRNA* families that have detectable *MIRNA*-related loci in their respective genomes (inner pie chart). The proportion of *MIRNA*-related loci in each class (inverted or direct duplication) forming transcripts that are predicted or validated targets of the miRNA formed from the duplication event (outer ring). **(F)** Proportion of *MIRNA*-related loci that are exons, introns, intergenic sequences, or MITEs (inner pie chart). Proportion of *MIRNA*-related loci forming transcripts that are predicted or validated targets of the miRNA formed from the duplication event (outer ring).

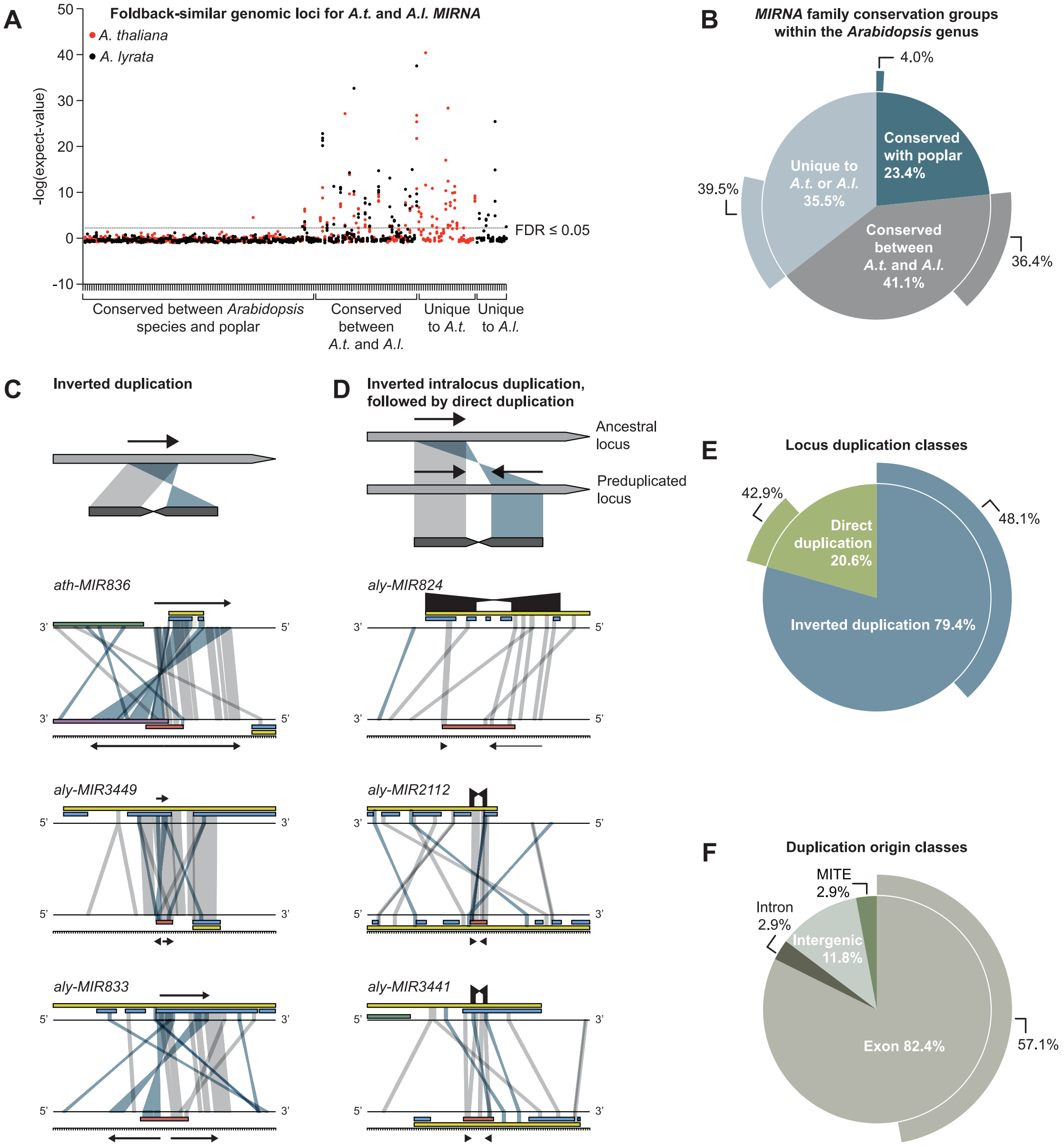


Figure 4.4. Identification of intragenomic loci with extended similarity to *MIRNA* genes.

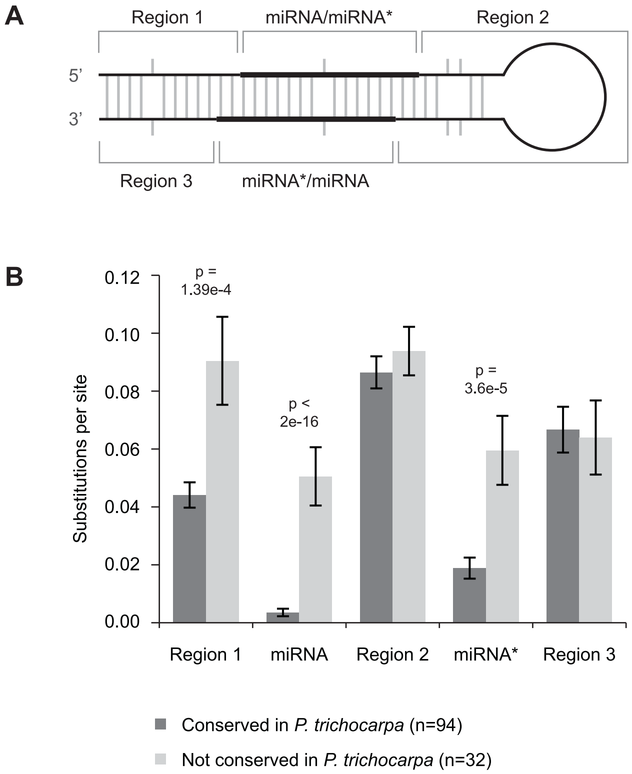


Figure 4.5. Sequence divergence within foldbacks of orthologous *A. thaliana* and *A. lyrata MIRNA*.

(A) *Arabidopsis MIRNA* foldbacks were divided into five regions. Note that the miRNA sequence can occur on either the 5’ or 3’ arm. **(B)** Sequence divergence of *Arabidopsis MIRNA*. Divergence was measured by adding nucleotide substitutions to insertion/deletion events for each *A. lyrata*/*A. thaliana* foldback alignment and dividing the sum by the regional sequence length. *Arabidopsis MIRNA* families were grouped into conservation groups based on whether or not the *MIRNA* family was conserved in *P. trichocarpa*. Listed P values are from pairwise permutation tests. Standard error bars are shown.

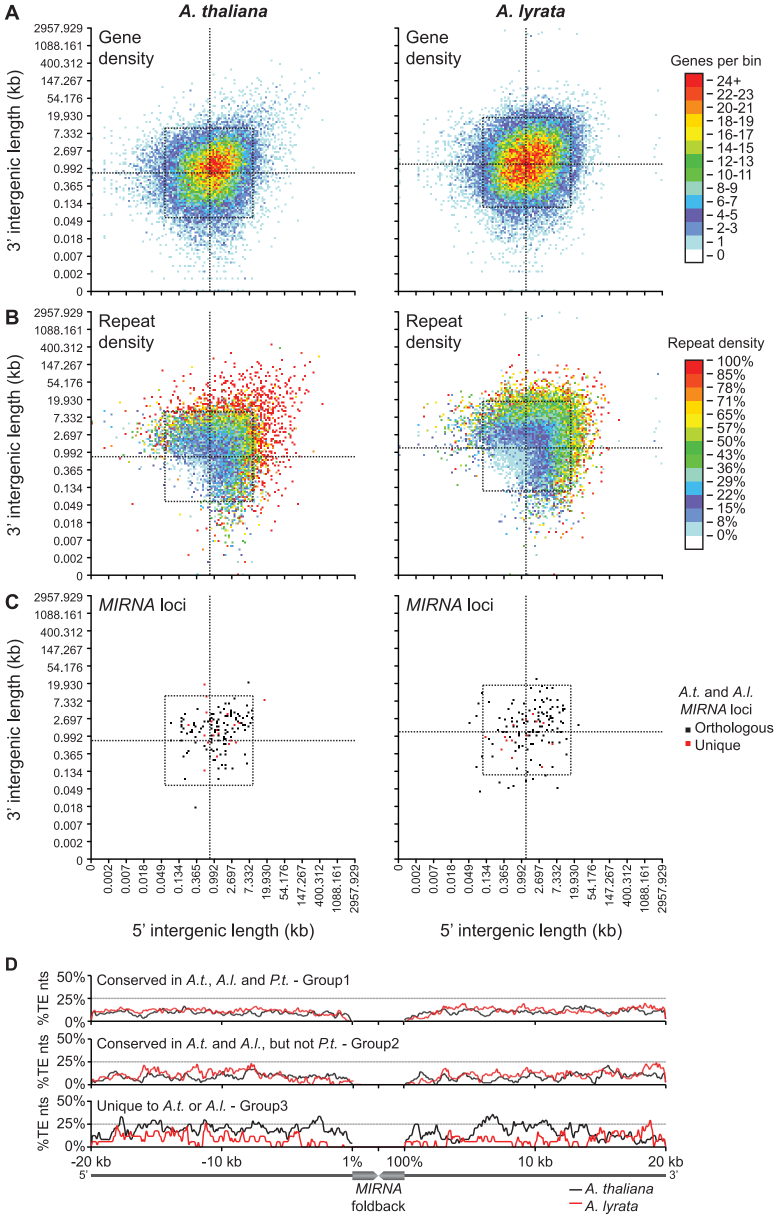


Figure 4.6. Gene and repeat density in the *A. thaliana* and *A. lyrata* genomes.
In **(A)** to **(C)**, annotated genes in *A. thaliana* and *A. lyrata* were placed in two-dimensional bins based on the length of their 5' and 3' intergenic lengths. **(A)** Gene density. The number of genes per bin is color coded. **(B)** Repeat density. The percentage of total nucleotides in the 5' and 3' intergenic regions of the genes in **(A)** occupied by transposable elements and repeat sequences is color coded. **(C)** *MIRNA* loci. Bins that contain *MIRNA* that have orthologs in *A. thaliana* and *A. lyrata* are colored black. Bins that contain unique *A. thaliana* or *A. lyrata* *MIRNA* are colored red. **(D)** Percentage of positions occupied by transposable elements in 20,000-nucleotide segments upstream and downstream of *A. thaliana* and *A. lyrata MIRNA*. Percentages were calculated in scrolling windows (100-nucleotide windows, 20-nucleotide scroll). *MIRNA* foldbacks are plotted on a relative scale.

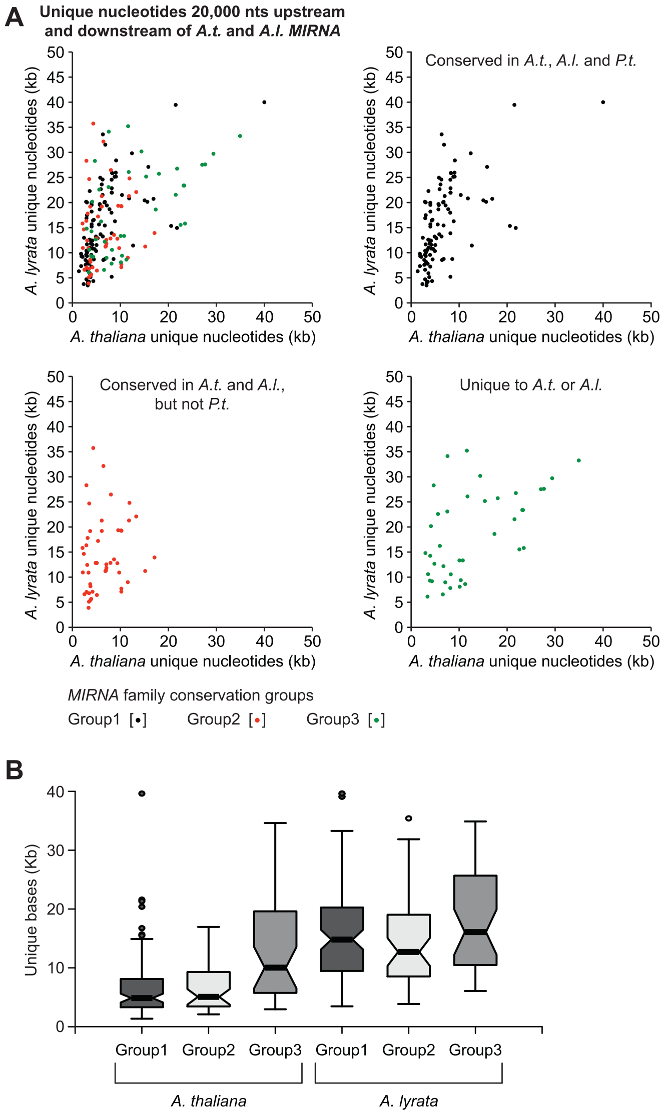


Figure 4.7. The genomic context of *A. thaliana* and *A. lyrata MIRNA*.
(A) The number of unique positions within 20,000 nucleotides upstream and downstream of *MIRNA* genes in *A. thaliana* and *A. lyrata* for which orthologous loci were identified. Scatterplots compare unique nucleotides in *A. thaliana* to *A. lyrata* for *MIRNA* families conserved to *P. trichocarpa* (black dots, Group 1), conserved only to *A. thaliana* and *A. lyrata* (red dots, Group 2), or *MIRNA* families that are unique to *A. thaliana* or *A. lyrata* (green dots, Group 3). The top left plot shows all three conservation groups. **(B)** Notched boxplots of unique positions shown in **(A)**. Notches mark the 95% confidence interval of the median for each group.

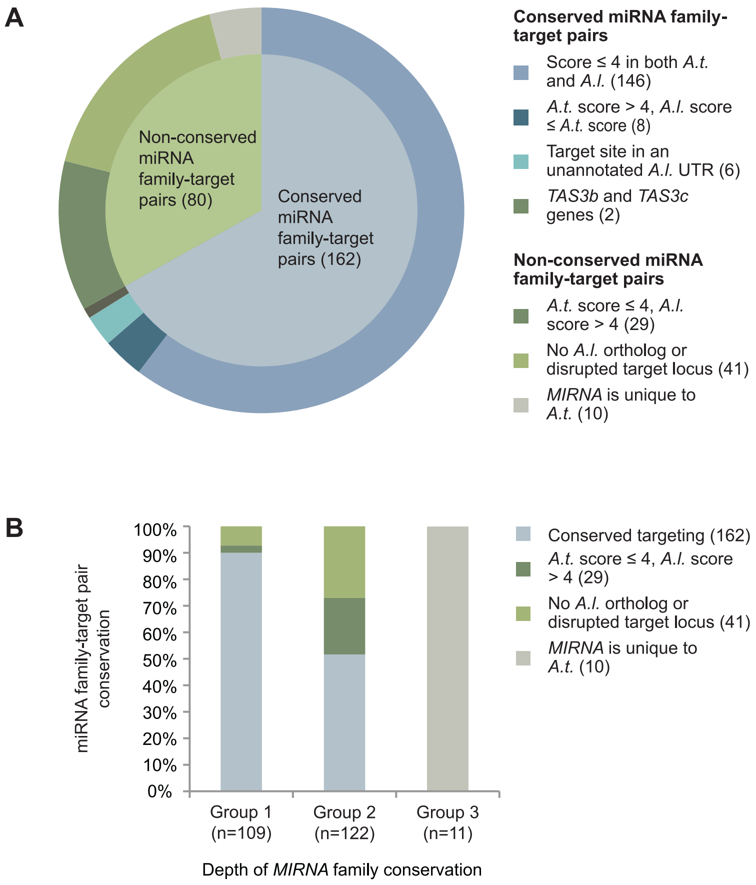


Figure 4.8. Conservation of *A. thaliana* miRNA-target pairs in *A. lyrata*.
(A) Conserved and nonconserved *A. thaliana* miRNA-target pairs in *A. lyrata* (inner pie chart). Numbers in parentheses indicate pair counts for each category. The outer ring shows criteria for inclusion into a conservation category. **(B)** Conservation or degradation of *A. thaliana* miRNA-target pairs. Pairs from **(A)** were binned into those present in *P. trichocarpa*, *A. lyrata*, and *A. thaliana* (Group 1), those present in *A. lyrata* and *A. thaliana* but not *P. trichocarpa* (Group 2), or those unique to *A. thaliana* (Group 3). The portion of conserved targeting, degraded targeting, and nonconserved targeting within each group is plotted. Note that degraded targeting can occur by substitutions in either the miRNA or target sequence.

SUPPLEMENTAL DATA

The following materials are available in the online version of this article (http://www.plantcell.org).

Supplemental Figure 3. *A. thaliana*, *A. lyrata*, and *C. rubella MIRNA* foldback sequences.

Supplemental Figure 4. Locus conservation around all *A. thaliana* and *A. lyrata MIRNA*s.

Supplemental Data Set 1A. *A. lyrata MIRNA* genes.

Supplemental Data Set 1B. *C. rubella MIRNA* genes.

Supplemental Data Set 1C. *MIRNA*-related loci in *A. thaliana* and *A. lyrata*.

Supplemental Data Set 1D. *A. thaliana* family miRNA-target pair conservation in *A. lyrata*.

Supplemental Data Set 1E. *A. lyrata* GENSCAN gene models.

Table S4.1. Small RNA library statistics

			Total small RNA		Mapped small RNA (15-30 nts)		
Species	Tissue	Method[a]	Reads	Unique	Unique	Reads	Genome hits
A. lyrata	Flowers (stage 1-12)	454 pyrosequencing	242,641	46,553	24,631	185,683	288,509
A. lyrata	Flowers (stage 1-12)	454 pyrosequencing	260,883	149,669	105,438	199,408	1,255,055
A. lyrata	Seedlings (14 days)	454 pyrosequencing	41,353	13,540	8,287	30,379	83,451
A. lyrata	Seedlings (14 days)	454 pyrosequencing	171,659	49,860	30,547	123,634	411,600
A. lyrata	Rosette leaves (vegetative)	Illumina SBS	5,970,967	2,229,739	889,740	4,036,969	9,551,535
A. lyrata	Flowers (stage 1-12)	Illumina SBS	6,406,534	2,717,420	1,636,295	4,876,856	14,426,190
A. lyrata	Flowers (stage 1-12)	Illumina SBS	5,427,123	2,181,982	1,369,093	4,229,434	11,868,305
		Totals:	18,521,160	6,560,772	3,360,832	13,682,363	28,054,366
C. rubella	Flowers (stage 1-12)	454 pyrosequencing	331,010	81,832	35,308	185,339	5,221,126
C. rubella	Seedlings (14 days)	454 pyrosequencing	191,996	83,062	47,000	124,377	10,663,189
C. rubella	Seedlings (5 days) 6 hr mock	454 pyrosequencing	93,855	9,393	2,444	52,372	864,995
C. rubella	Seedlings (5 days) 6 hr 150 mM NaCl	454 pyrosequencing	306,425	71,982	30,355	186,791	9,070,137
		Totals:	923,286	231,196	115,107	548,879	19,517,826
A. thaliana	Total aerial (bolting, flowering) 21 days	Illumina SBS	6,293,256	1,992,169	741,843	3,903,749	2,578,015

[a] 454 pyrosequencing (454 Life Sciences, http://www.454.com). Illumina SBS (Illumina, Inc., http://www.illumina.com).

Table S4.2. Summary statistics for *A. lyrata* and *A. thaliana* annotated repeat features

	A. thaliana		***A. lyrata***	
Feature	**Count**	**Total size (bp)**	**Count**	**Total size (bp)**
Transposon	10184	6906214	21845	7902856
Composite	157	56068	535	131984
Helitron	12945	7548301	11960	3567995
LTR Retrotransposon	5962	8894632	20169	18362274
Non-LTR Retrotransposon	1985	1379023	6496	4502410
Satellite/Centromeric	838	813818	2727	1129963
Inverted repeat	68111	7711655	102516	11990169
Tandem repeat	136105	7666949	226023	14506279
Totals:	236287	40976660	392271	62093930

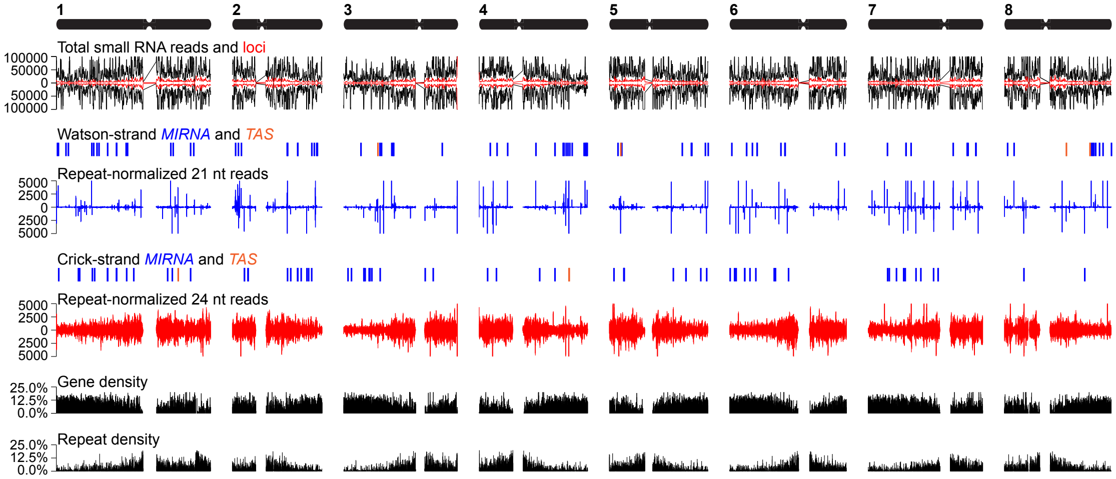


Figure S4.1. Distribution of small RNA, genes and repeat elements in the *A. lyrata* genome.

A. lyrata chromosomes 1-8 are depicted as ideograms at the top of the figure. Total small RNA reads and loci graphs plot total small RNA reads (black lines) and total small RNA loci (red lines). Repeat-normalized 21- and 24-nucleotide reads histograms plot 21- or 24-nucleotide small RNA reads normalized for repeated matches to the genome. Watson- and Crick-strand *MIRNA* and *TAS* plots mark the positions of *MIRNA* (blue) and *TAS* (green) genes on the Watson or Crick strand, respectively. Gene and repeat density histograms plot the percentage of nts per window occupied by genes (exons + introns) or repeats (transposons, retrotransposons, and centromeric repeats). Histogram counts for total reads, total loci, repeat-normalized 21-nucleotide reads, repeat-normalized 24-nucleotide reads, gene density and repeat density were computed using the scrolling window method (window size = 100000 nucleotides, scroll length = 20000 nucleotides).

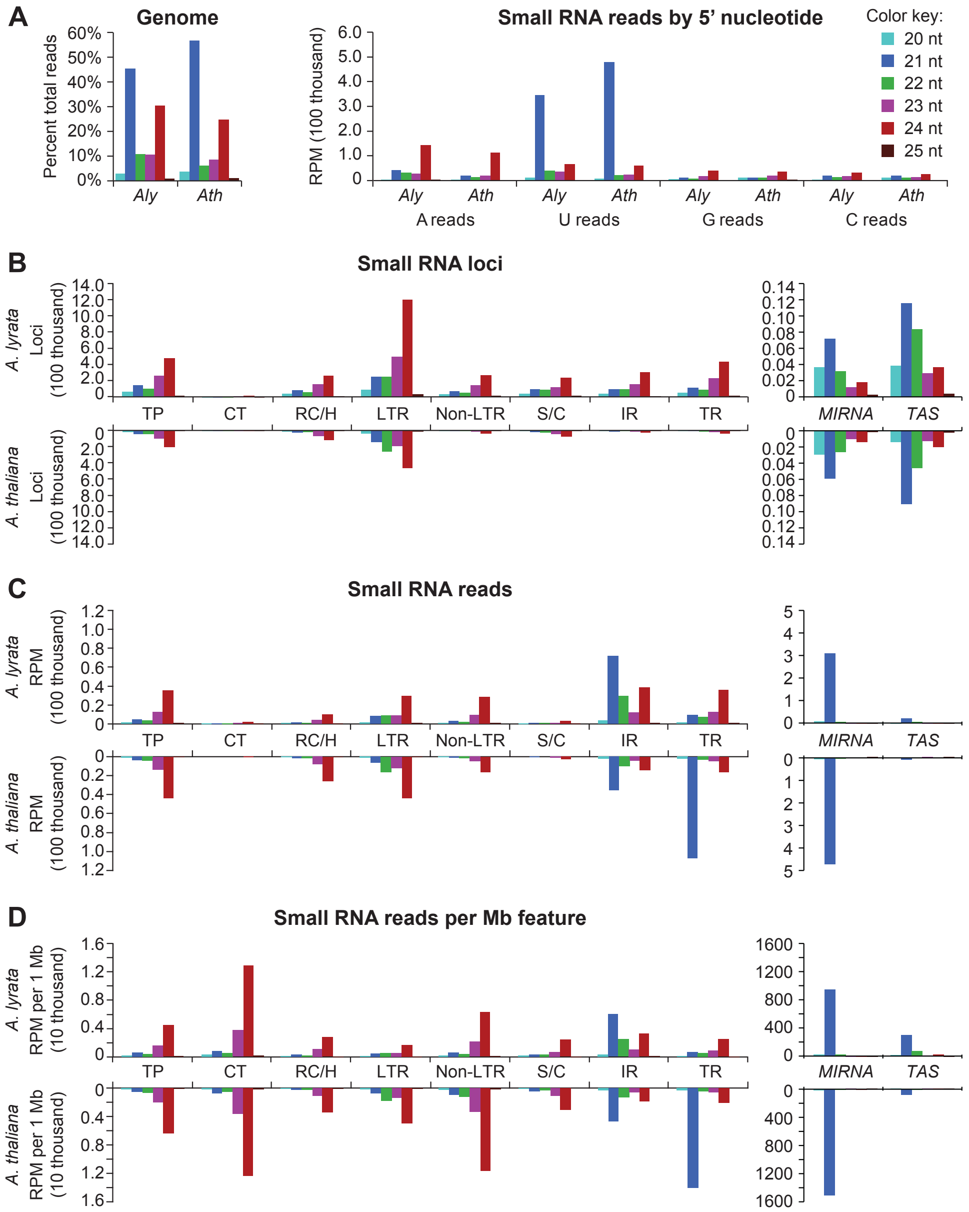


Figure S4.2. Basic profile of small RNA in *A. lyrata* and *A. thaliana*.
(A-D) Distribution of 20-25 nucleotide small RNA for *A. lyrata* and *A. thaliana*. **(A)** Percent of total normalized small RNA reads in each size class (left) and distribution of normalized reads for small RNA with a 5' A, U, G or C (right). **(B-D)** Distribution of small RNA in transposons (TP), composite transposons (CT), helitrons (RC/H), LTR retrotransposons (LTR), non-LTR retrotransposons (Non-LTR), satellite and centromeric repeats (S/C), inverted repeats (IR), and tandem repeats (TR) by loci **(B)**, repeat-normalized RPM **(C)**, and repeat-normalized RPM per 1 Mb of feature **(D)**. *A. lyrata* data are plotted up and *A. thaliana* data are plotted down in **(B-D)**. In all panels small RNA sizes are colored coded [see color key in panel **(A)**, right].

General Conclusions

Noah Fahlgren, Josh T. Cuperus and James C. Carrington

In preparation for submission to *The Plant Cell*
American Society of Plant Biologists
15501 Monona Drive
Rockville, Maryland 20855-2768

EVOLUTION OF *MIRNA* GENES

Old and young plant *MIRNA* families

The majority of miRNAs that were discovered in early studies of *Arabidopsis thaliana* and *Oryza sativa* were from families that are conserved between the two species (Llave et al., 2002; Mette et al., 2002; Park et al., 2002; Reinhart et al., 2002; Jones-Rhoades and Bartel, 2004; Sunkar and Zhu, 2004). miRNA from these loci were detected by low-throughput cloning and sequencing, or by computational predictions of foldbacks and complementary target sequences. These techniques resulted in a bias for detecting miRNA from abundantly expressed and ancient families (Axtell and Bartel, 2005). Combined with numerous datasets from high-throughput sequencing approaches, eight *MIRNA* families have been identified in the common ancestor of all embryophytes (Figure 5.1). The *MIR396* family was present in the common ancestor of all tracheophytes (vascular plants), while the *MIR397* and *MIR398* families were acquired in the common ancestor of all spermatophytes (seed plants; Figure 5.1). Ten families are present in all angiosperm lineages, while all other families display more restricted taxonomic distributions. *MIR828*, *MIR2111* and *MIR403* are eudicot-specific families, though *MIR403* may be restricted to core eudicot lineages (Figure 5.1). *MIR472* is present in all core rosids, while *MIR857* may be restricted to at least a subset of the eurosid II clade. At least nine families likely arose in the monocot lineage. Four of these are found in all three grass families, and five were lost in one lineage (Figure 5.1). Two additional family patterns suggest lineage-specific losses. *MIR529* and *MIR536* were present in the common ancestor of embryophytes but were lost in the common ancestors of eudicots and tracheophytes, respectively (Figure 5.1). No *MIRNA* families found in embryophyte plants have been found in the unicellular green alga *Chlamydomonas reinhardtii* (Molnar et al., 2007; Zhao et al., 2007). Despite this, two *MIRNA* annotated families (*MIR854* and *MIR855*) were reported as conserved between the plant and animal lineages (Arteaga-Vazquez et al., 2006). However, loci from these families lie in retrotransposons. Analysis of deep sequencing data at these loci in *Arabidopsis* have failed to validate these as bona fide *MIRNA*, but rather to be loci that spawn heterogeneous, RDR2/DCL3-dependent, 24-nucleotide siRNA (Lu et al., 2006; Rajagopalan et al., 2006; Fahlgren et al., 2007; Kasschau et al., 2007; Fahlgren et al., 2010; Ma et al., 2010). Current evidence, therefore, suggests that known *MIRNA* families in plants and animals arose independently.

Deep sequencing of *A. thaliana* small RNA populations revealed that, in addition to *MIRNA* that are conserved between at least two families (Figure 5.1), there are at least four times as many families that are not obviously conserved outside of Brassicaceae (Lu et al., 2006; Rajagopalan et al., 2006; Fahlgren et al., 2007; Fahlgren et al., 2010; Ma et al., 2010).

A high proportion of species-specific or non-conserved *MIRNA* was also observed in *Physcomitrella patens* and *Selaginella moellendorffii* (Axtell et al., 2007), *O. sativa* (Heisel et al., 2008; Lu et al., 2008a; Sunkar et al., 2008; Zhu et al., 2008), *Medicago truncatula* (Szittya et al., 2008; Lelandais-Briere et al., 2009) and *Glycine max* (Subramanian et al., 2008). Data from additional plant genomes indicate that the majority of these are restricted to species or other sub-family lineages (http://www.mirbase.org; Griffiths-Jones et al., 2008). Given that a large number of *MIRNA* families are species-specific or restricted to closely related species, it is reasonable to suggest that plants harbor relatively large numbers of recently spawned *MIRNA* loci.

Origins and evolution of new *MIRNA* genes

How are new MIRNA formed?

Early sequence similarity searches revealed that the foldback arms of *ath-MIR161*, *ath-MIR163* and *ath-MIR822* (previously referred to as ASRP1729) shared extended similarity (outside of the miRNA/miRNA* regions) with their target genes (Allen et al., 2004). The foldback arms aligned in an inverted orientation, with the miRNA arm inverted relative to the target gene sequence. These data suggested that inverted duplication events could form self-complementary regions with the potential to spawn *MIRNA* genes. Initial duplication events would result in loci with perfect or near-perfect self-complementarity and produce siRNA (Allen et al., 2004). Indeed, it was demonstrated experimentally that the *ath-MIR822*, *ath-MIR839* and *ath-MIR869* foldbacks are processed by DCL4, rather than DCL1, into miRNA-like siRNA (Rajagopalan et al., 2006; Ben Amor et al., 2009). These *MIRNA* genes formed by an inverted duplication of a DC1 domain gene (*MIR822*), a P-glycoprotein gene (*MIR839*) or a SU(VAR)3-9 homolog gene (*MIR869*) (Allen et al., 2004; Fahlgren et al., 2007; Fahlgren et al., 2010), and thus represent good examples of young, transitional *MIRNA*. Over time, accumulation of mutations in the foldback arms could result in an affinity for the miRNA biogenesis machinery, and decreased similarity between the extended foldback arms and the locus of origin (Allen et al., 2004). Coupled with the co-evolution of one or more target transcripts, a productive miRNA-target node may arise and, in rare instances, be incorporated into new or existing regulatory networks.

Further evidence for the recent origin of some *MIRNA* was found by global analysis of genomic or transcript sequences with extensive similarity to *MIRNA* loci in *A. thaliana*, *A. lyrata* and *P. patens* (Rajagopalan et al., 2006; Axtell et al., 2007; Fahlgren et al., 2007; de Felippes et al., 2008; Fahlgren et al., 2010). 27 and 18 *MIRNA* loci in *A. thaliana* and *A. lyrata*, respectively, contained sequence similarity with a putative locus of origin elsewhere in their respective genome, with a large proportion of these in *A. thaliana* originating through duplications of transcribed, protein-coding gene sequences (Rajagopalan et al., 2006; Axtell et al., 2007; Fahlgren et al., 2007; de Felippes et al., 2008; Fahlgren et al., 2010). However, other configurations were also detected. For some *MIRNA*, both arms of the foldback align to the putative foldback-originating locus in the same orientation, or align to separate regions rather than one, suggesting that a duplication event occurred at the originating locus before the duplication event that formed the *MIRNA* (Rajagopalan et al., 2006; Fahlgren et al., 2010).

Despite the large number of young *A. thaliana* and *A. lyrata MIRNA* with sequence identity to putative loci of origin, over half possess no identifiable *MIRNA*-related locus (de Felippes et al., 2008; Fahlgren et al., 2010). Where did these *MIRNA* come from? Conceivably, some of the originating loci were lost, or rapidly diverged. For example *MIR161*, *MIR472* and *MIR822* had significant matches with other loci in *A. thaliana*, but not in *A. lyrata*, while *MIR859* had a significant match in *A. lyrata*, but not in *A. thaliana* (Fahlgren et al., 2010). On the other hand, any inverted duplication in a plant genome could be the starting point for a new *MIRNA*. Jones-Rhoades and Bartel (2004) identified 133,864 and 410,167 imperfect inverted repeats in the genomes of *A. thaliana* and *O. sativa*, respectively. One source of inverted repeats are non-autonomous transposons that contain flanking terminal inverted repeats, but that lack the internal sequences encoding genes required for transposition (miniature inverted-repeat transposable element; MITE) (Bureau and Wessler, 1992, 1994). A MITE locus may have been the source of *ath-MIR1888*, which requires DCL1 for biogenesis, but which also has similarity to a number of small inverted repeats that generate DCL3-dependent 24-nucleotide siRNA (German et al., 2008). Several additional *A. thaliana* and *O. sativa* miRNA were proposed to be derived from transposable elements (Piriyapongsa and Jordan, 2008), although many of these are not bona fide *MIRNA* (see miR854 and miR855 above; Axtell and Bowman, 2008; Meyers et al., 2008). While the transposon origin model is plausible, transitionary (young) loci may be difficult to identify due to the intersection of miRNA and siRNA pathways at these loci (Chellappan et al., 2010).

A final possibility for the origin of young *MIRNA* that lack detectable *MIRNA*-related loci is that they were not formed by duplication events. Instead, new *MIRNA* genes could form through the accumulation of mutations within inverted repeats (de Felippes et al., 2008). If expressed, selection could conceivably act on mutations that modify miRNA processing efficiency and alter affinity of the de novo-generated miRNA for a target if there were effects on existing regulatory networks.

What is the rate that MIRNA are formed or lost?

The abundance of species- or lineage-specific *MIRNA* families suggests that *MIRNA* genes are born and lost at a high frequency. The recent high-quality assembly of the *A. lyrata* genome (http://genome.jgi-psf.org/Araly1/Araly1.home.html) allowed for a thorough analysis of the shared and unique *MIRNA* contents of the *A. thaliana* and *A. lyrata* genomes (Fahlgren et al., 2010; Ma et al., 2010), providing the basis for the first *MIRNA* flux estimates. There are 102 and 116 identified *MIRNA* families in *A. thaliana* and *A. lyrata*, respectively, with 78 families shared between the two (http://www.mirbase.org; Griffiths-Jones et al., 2008; Fahlgren et al., 2010; Ma et al., 2010). Although the majority of *MIRNA* families are conserved between the two species, 24-33% of the families were gained or lost by *A. thaliana* or *A. lyrata* since they diverged approximately 10 million years ago (Koch et al., 2000; Wright et al., 2002; Ossowski et al., 2010). Additionally, preliminary mining of the *Capsella rubella* genome and small RNA sequencing data identified 43 and 42 *MIRNA* families conserved with *A. thaliana* or *A. lyrata*, respectively (Fahlgren et al., 2010). Given that *C. rubella* diverged from the *Arabidopsis* lineage ~20 million years ago (Koch et al., 2000; Wright et al., 2002; Ossowski et al., 2010), and allowing for error identifying *MIRNA* in the incomplete *C. rubella* genome, the net rate of flux (birth-death) for *MIRNA* in the *Arabidopsis* lineage was estimated to be from 1.2—3.3 genes per million years (Fahlgren et al., 2010). This estimate overlaps the 0.8—1.6 genes per million years estimated for the *Drosophila* lineage *MIRNA* flux rate (Berezikov et al., 2010).

Is the formation of new MIRNA a selective or neutral process?

Young *MIRNA* genes formed by duplications of transcribed regions could yield miRNA, or siRNA, that have the potential to perturb existing regulatory networks. However, evidence from several plant species suggests that the vast majority of young miRNAs have few, if any, functions. First, whereas approximately 45% of *A. thaliana* targets for conserved miRNAs increase in abundance in multiple mutants with miRNA pathway defects (*hyl1*, *hst*, *dcl1*, *hen1* and *ago1*; Allen et al., 2005; Ronemus et al., 2006), accumulation of target transcripts for non-conserved miRNAs are largely unaffected by miRNA biogenesis mutants, suggesting that most young miRNA are not integrated into regulatory networks (Fahlgren et al., 2007). Target conservation patterns between *A. thaliana* and *A. lyrata* also support a neutral regulatory role for most young miRNA in these species. For *MIRNA* families conserved with non-Brassicaceae species, prediction and validation of miRNA targets was highly reliable and consistent between species (Ma et al., 2010). In contrast, target prediction and validation for miRNA conserved between *A. thaliana* and *A. lyrata*, but not other plant families, was limited and highly species-specific (Ma et al., 2010). This supports the earlier finding that most young miRNA in *A. thaliana* have few bona fide targets (Rajagopalan et al., 2006; Fahlgren et al., 2007).

Second, the abundance of conserved miRNA, as a group, is substantially higher than the abundance of young miRNAs. Deeply conserved miRNA are consistently sequenced at a higher frequency and cumulatively make up a larger portion of the total reads from all *MIRNA* families (Rajagopalan et al., 2006; Axtell, 2008; Ma et al., 2010). In contrast, *Arabidopsis*-specific miRNA are generally less abundant and often expressed in only one of the two species (Ma et al., 2010). Data from a recent study that profiled *MIRNA* transcripts in miRNA biogenesis mutants and under stress conditions suggests that this bias is also present at the primary *MIRNA* transcript level (Laubinger et al., 2010). Over 90% of the families of *MIRNA* conserved between plant families had at least one member that was detectable at moderate to high levels, while less than half of the Brassicaceae-specific *MIRNA* families accumulated to the same levels (Laubinger et al., 2010). This could be the result of highly specific expression patterns in cells, tissues or conditions that have not been studied. Alternatively, the majority of young *MIRNA* may lack regulatory elements that confer robust expression. The duplication that formed *MIR163* was found to contain part of the promoter sequence of the originating target gene family (Wang et al., 2006), but similar patterns have not been reported for other young *MIRNA*. In addition to expression, young *MIRNA* are also processed less precisely. Those apparent transitional *MIRNA* families with transcripts that are processed by DCL4 exhibit the most striking examples of imprecise processing (see above; Rajagopalan et al., 2006), but processing precision is generally low for Brassicaceae-restricted *MIRNA* transcripts (Vazquez et al., 2008; Ma et al., 2010). Weak expression coupled with variable processing may partly explain the lack of experimental support for the majority of young *MIRNA*.

Finally, nucleotide divergence patterns between orthologous *MIRNA* in *A. thaliana* and *A. lyrata* are consistent with neutral evolution. For *Arabidopsis MIRNA* families conserved with non-Brassicaceae species, nucleotide divergence was highest in the loop and loop-distal stem regions, and lowest in the miRNA and miRNA* regions (Ehrenreich and Purugganan, 2008; Warthmann et al., 2008; Fahlgren et al., 2010; Ma et al., 2010). Low divergence in the miRNA region likely reflects purifying selection to maintain complementarity between the mature miRNA and target RNAs, while the low divergence in the miRNA* region likely reflects purifying selection to maintain base pairing with the constrained miRNA. In contrast, nucleotide divergence was much more uniform across the foldbacks of *MIRNA* not conserved with non-Brassicaceae species (Fahlgren et al., 2010; Ma et al., 2010). Although the pairwise differences were somewhat lower in the miRNA and miRNA* regions, divergence was significantly higher than that found for the deeply conserved *MIRNA* families (Fahlgren et al., 2010; Ma et al., 2010). While some young MIRNA may experience strong selection, these data suggest that selection is weak or neutral for the majority of these loci.

In contrast to deeply conserved *MIRNA* families, most young *MIRNA* are weakly expressed, processed imprecisely, more divergent and tend to lack targets, suggesting that they may be evolving neutrally. The lack of verifiable targets may be due to the low accumulation and imprecise generation of mature miRNA. For instance, imprecise excision of the mature miRNA could result in functionally variable, or inert, miRNA as the miRNA sequence is important for selection of the miRNA over the miRNA* and sorting into functional AGO complexes (Mi et al., 2008; Montgomery et al., 2008a; Eamens et al., 2009; Ebhardt et al., 2010). Alternatively, young miRNA could function primarily through non-degradative mechanisms that cannot be assayed by measuring target transcript abundance (Brodersen et al., 2008; Lanet et al., 2009). While this cannot be ruled out, it is difficult to explain why such a preference would exist. Additionally, the higher degree of nucleotide divergence observed for young *MIRNA* orthologs and the observed birth-death rate for *Arabidopsis MIRNA* supports the idea that the most are neutrally evolving, evolutionarily transient loci. New loci would have little detrimental impact on existing regulatory networks, but could be sources of novel regulatory variation that are captured into networks on relatively rare occasions. The Brassicaceae-specific *MIR824* may be such an example. miR824-guided cleavage of *AGAMOUS-LIKE16* transcripts functions in a stomatal patterning developmental network (Kutter et al., 2007), and population variation at the *ath-MIR824* locus shows significant departure from neutrality (de Meaux et al., 2008).

CONCLUSIONS AND FUTURE PERSPECTIVES

The rapidly growing collection of plant genomes, and advances in high-throughput small RNA sequencing has enhanced our understanding of the evolution of *MIRNA* genes and their functions. Improvements in accuracy, depth-of-coverage and lower cost has made high-throughput sequencing one of the most prevalent techniques used to study whole small RNA populations. For these same reasons, small RNA profiling studies should incorporate the use of replicates with statistical analysis. However, further research is needed to determine the optimal statistical models for various kinds of small RNA profiling experiments. Additionally, next-generation alignment tools like CASHX, BOWTIE and SOAP have largely overcome initial problems aligning massive numbers of sequencing reads to a reference genome (Fahlgren et al., 2009; Langmead et al., 2009; Li et al., 2009), but tools for storage and retrieval of data that scale well with hundreds of millions or even billions of reads are still lacking.

Several *MIRNA* families are ancient and were found in the common ancestor of land plants. However, improved taxonomic coverage has revealed that many *MIRNA* families are lineage-restricted or species-specific. Deeply conserved *MIRNA* families are integral components of important regulatory networks that orchestrate developmental and stress responses. It is tempting to hypothesize that some well-conserved, lineage-restricted *MIRNA* were important for species-diversification. In contrast, most young *MIRNA*, those found only among closely related species, seem to be selectively neutral. In most cases, these families are lost over short evolutionary periods, but occasionally are maintained by selection. The work presented here focused on the acquisition and subsequent evolution of new *MIRNA*, but evolution of established *MIRNA* families is also important. Established families evolve by duplication, or deletion, followed by sub or neofunctionalization (Maher et al., 2006; Warthmann et al., 2008). Paralogous *MIRNA* genes can diversify regulatory elements, resulting in spatial or temporal partitioning of pri-miRNA transcripts (Voinnet, 2009). Our understanding of *MIRNA* evolution and function has rapidly improved through genome-wide analysis, but in many cases the functions (redundant or not) of individual *MIRNA* genes remains unknown or explored only in the experimental model *A. thaliana*. Already, analysis of small RNA populations in *O. sativa* and *Brachypodium distachyon* discovered clusters of miRNA-triggered 24-nucleotide secondary siRNA, a population that is not found in *A. thaliana*, demonstrating a need for studies of small RNA populations in other plant systems (Johnson et al., 2009; International Brachypodium Initiative, 2010). Additionally, analysis of lineage-restricted *MIRNA* among groups of closely related plants in diverse lineages will be used to test the evolutionary hypotheses generated by comparing *A. thaliana*, *A. lyrata* and *C. rubella*. As *MIRNA* are identified in new species, the necessity for good annotation will be critical for evolutionary studies. Care should be taken to annotate new *MIRNA* using community standards (Meyers et al., 2008).

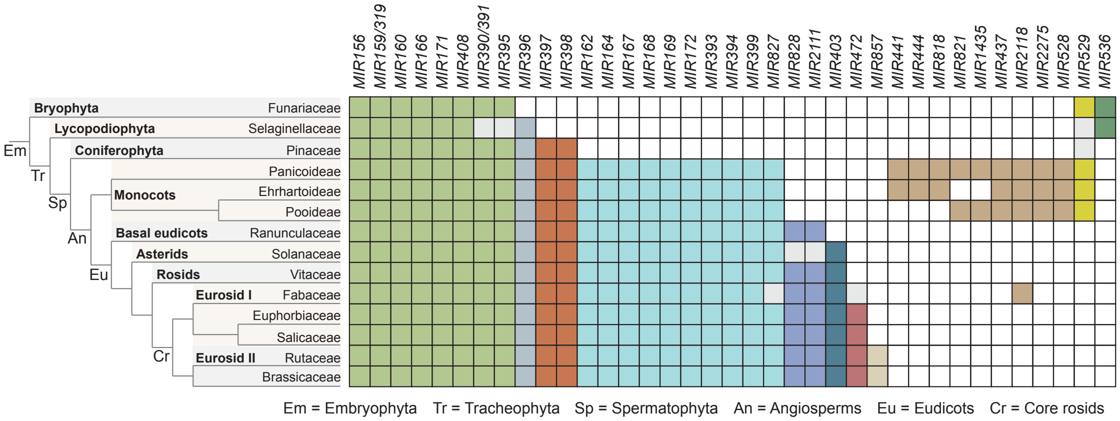


Figure 5.1. Deeply conserved *MIRNA* families.
MIRNA families (columns) conserved between plant families (rows) for plant species represented in miRBase release 16 (Griffiths-Jones et al., 2008). Species- and family-specific *MIRNA* were omitted. Boxes are highlighted if a *MIRNA* family was identified in at least one species for each of the plant families listed, or if the *MIRNA* family could be identified in the National Center for Biotechnology Information (NCBI) EST or whole-genome shotgun reads databases by BLAST (Altschul et al., 1997). Matches were identified as *MIRNA* if the sequences containing the putative mature miRNA could be folded, using RNAfold (Hofacker, 2003), into a stem-loop structure containing the miRNA in the stem. Plant families that may have lost a *MIRNA* family, or where the family could not confidently be identified, are shaded gray. Boxes are highlighted different colors and grouped based on inferred taxonomic range for each *MIRNA* family.

Bibliography

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.

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

Allison, D.B. (2006). DNA microarrays and related genomics techniques: designs, analysis, and interpretation of experiments. Chapman and Hall/CRC, Boca Raton, F.L.

Allison, D.B., Cui, X., Page, G.P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet **7,** 55-65.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215,** 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25,** 3389-3402.

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

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796-815.

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. (2008). Evolution of microRNAs and their targets: are all microRNAs biologically relevant? Biochim Biophys Acta **1779,** 725-734.

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.

Axtell, M.J., Snyder, J.A., and Bartel, D.P. (2007). Common functions for diverse small RNAs of land plants. Plant Cell **19,** 1750-1769.

Axtell, M.J., Jan, C., Rajagopalan, R., and Bartel, D.P. (2006). A two-hit trigger for siRNA biogenesis in plants. Cell **127,** 565-577.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res **33,** D562-566.

Batista, P.J., Ruby, J.G., Claycomb, J.M., Chiang, R., Fahlgren, N., Kasschau, K.D., Chaves, D.A., Gu, W., Vasale, J.J., Duan, S., Conte, D., Jr., Luo, S., Schroth, G.P., Carrington, J.C., Bartel, D.P., and Mello, C.C. (2008). PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. Mol Cell **31,** 67-78.

Baulcombe, D. (2004). RNA silencing in plants. Nature **431,** 356-363.

Baumberger, N., and Baulcombe, D.C. (2005). *Arabidopsis* ARGONAUTE1 is an RNA slicer that selectively recruits microRNAs and short interfering RNAs. Proc Natl Acad Sci USA **102,** 11928-11933.

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

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27,** 573-580.

Bentley, D.R. (2006). Whole-genome re-sequencing. Curr Opin Genet Dev **16,** 545-552.

Berezikov, E., Liu, N., Flynt, A.S., Hodges, E., Rooks, M., Hannon, G.J., and Lai, E.C. (2010). Evolutionary flux of canonical microRNAs and mirtrons in *Drosophila*. Nat Genet **42,** 6-9; author reply 9-10.

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.

Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol **13,** 1097-1101.

Bray, N., and Pachter, L. (2004). MAVID: constrained ancestral alignment of multiple sequences. Genome Res **14,** 693-699.

Bray, N., Dubchak, I., and Pachter, L. (2003). AVID: A global alignment program. Genome Res **13,** 97-102.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell **128,** 1089-1103.

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.

Bureau, T.E., and Wessler, S.R. (1992). Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell **4,** 1283-1294.

Bureau, T.E., and Wessler, S.R. (1994). Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell **6,** 907-916.

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

Chan, S.W., Henderson, I.R., and Jacobsen, S.E. (2005). Gardening the genome: DNA methylation in *Arabidopsis thaliana*. Nat Rev Genet **6,** 351-360.

Chan, S.W., Zhang, X., Bernatavichute, Y.V., and Jacobsen, S.E. (2006). Two-step recruitment of RNA-directed DNA methylation to tandem repeats. PLoS Biol **4,** e363.

Chapman, E.J., and Carrington, J.C. (2007). Specialization and evolution of endogenous small RNA pathways. Nat Rev Genet **8,** 884-896.

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

Chen, H.M., Chen, L.T., Patel, K., Li, Y.H., Baulcombe, D.C., and Wu, S.H. (2010). 22-nucleotide RNAs trigger secondary siRNA biogenesis in plants. Proc Natl Acad Sci U S A **107,** 15269-15274.

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

Chen, X. (2010). Small RNAs - secrets and surprises of the genome. Plant J **61,** 941-958.

Chitwood, D.H., Nogueira, F.T., Howell, M.D., Montgomery, T.A., Carrington, J.C., and Timmermans, M.C. (2009). Pattern formation via small RNA mobility. Genes Dev **23,** 549-554.

Chu, C.Y., and Rana, T.M. (2006). Translation repression in human cells by microRNA-induced gene silencing requires RCK/p54. PLoS Biol **4,** e210.

Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T.T., Fu, G., Hinds, D.A., Chen, H., Frazer, K.A., Huson, D.H., Scholkopf, B., Nordborg, M., Ratsch, G., Ecker, J.R., and Weigel, D. (2007). Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science **317,** 338-342.

Cullen, B.R. (2009). Viral RNAs: lessons from the enemy. Cell **136,** 592-597.

Cuperus, J.T., Montgomery, T.A., Fahlgren, N., Burke, R.T., Townsend, T., Sullivan, C.M., and Carrington, J.C. (2010a). Identification of *MIR390a* precursor processing-defective mutants in *Arabidopsis* by direct genome sequencing. Proc Natl Acad Sci U S A **107,** 466-471.

Cuperus, J.T., Carbonell, A., Fahlgren, N., Garcia-Ruiz, H., Burke, R.T., Takeda, A., Sullivan, C.M., Gilbert, S.D., Montgomery, T.A., and Carrington, J.C. (2010b). Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in *Arabidopsis*. Nat Struct Mol Biol **17,** 997-1003.

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

de 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.

de Meaux, J., Hu, J.Y., Tartler, U., and Goebel, U. (2008). Structurally different alleles of the *ath-MIR824* microRNA precursor are maintained at high frequency in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **105,** 8994-8999.

Deleris, A., Gallego-Bartolome, J., Bao, J., Kasschau, K.D., Carrington, J.C., and Voinnet, O. (2006). Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. Science **313,** 68-71.

Dewey, C.N. (2007). Aligning multiple whole genomes with Mercator and MAVID. Methods Mol Biol **395,** 221-236.

Ding, S.W., and Voinnet, O. (2007). Antiviral immunity directed by small RNAs. Cell **130,** 413-426.

Dong, Z., Han, M.H., and Fedoroff, N. (2008). The RNA-binding proteins HYL1 and SE promote accurate in vitro processing of pri-miRNA by DCL1. Proc Natl Acad Sci U S A **105,** 9970-9975.

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

Eamens, A.L., Smith, N.A., Curtin, S.J., Wang, M.B., and Waterhouse, P.M. (2009). The *Arabidopsis thaliana* double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. RNA **15,** 2219-2235.

Ebhardt, H.A., Fedynak, A., and Fahlman, R.P. (2010). Naturally occurring variations in sequence length creates microRNA isoforms that differ in Argonaute effector complex specificity. Silence **1,** 12.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res **30,** 207-210.

Ehrenreich, I.M., and Purugganan, M.D. (2008). Sequence variation of microRNAs and their binding sites in *Arabidopsis*. Plant Physiol **146,** 1974-1982.

Faehnle, C.R., and Joshua-Tor, L. (2007). Argonautes confront new small RNAs. Curr Opin Chem Biol **11,** 569-577.

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

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.

Fan, J., Tam, P., Woude, G.V., and Ren, Y. (2004). Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. Proc Natl Acad Sci U S A **101,** 1135-1140.

Fang, Y., and Spector, D.L. (2007). Identification of nuclear dicing bodies containing proteins for microRNA biogenesis in living *Arabidopsis* plants. Curr Biol **17,** 818-823.

Farazi, T.A., Juranek, S.A., and Tuschl, T. (2008). The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. Development **135,** 1201-1214.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. (2005). The widespread impact of mammalian microRNAs on mRNA repression and evolution. Science **310,** 1817-1821.

Fujioka, Y., Utsumi, M., Ohba, Y., and Watanabe, Y. (2007). Location of a possible miRNA processing site in SmD3/SmB nuclear bodies in *Arabidopsis*. Plant Cell Physiol **48,** 1243-1253.

Gasciolli, V., Mallory, A.C., Bartel, D.P., and Vaucheret, H. (2005). Partially redundant functions of *Arabidopsis* DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. Curr Biol **15,** 1494-1500.

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.

Ghildiyal, M., Seitz, H., Horwich, M.D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E.L., Zapp, M.L., Weng, Z., and Zamore, P.D. (2008). Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. Science **320,** 1077-1081.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature **442,** 199-202.

Gregory, B.D., O'Malley, R.C., Lister, R., Urich, M.A., Tonti-Filippini, J., Chen, H., Millar, A.H., and Ecker, J.R. (2008). A link between RNA metabolism and silencing affecting *Arabidopsis* development. Dev Cell **14,** 854-866.

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

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., Chiang, H.R., King, N., Degnan, B.M., Rokhsar, D.S., and Bartel, D.P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature **455,** 1193-1197.

Guddeti, S., Zhang, D.C., Li, A.L., Leseberg, C.H., Kang, H., Li, X.G., Zhai, W.X., Johns, M.A., and Mao, L. (2005). Molecular evolution of the rice miR395 gene family. Cell Res **15,** 631-638.

Haag, J.R., Pontes, O., and Pikaard, C.S. (2009). Metal A and metal B sites of nuclear RNA polymerases Pol IV and Pol V are required for siRNA-dependent DNA methylation and gene silencing. PLoS One **4,** e4110.

Haas, B.J., Kamoun, S., Zody, M.C., Jiang, R.H., Handsaker, R.E., Cano, L.M., Grabherr, M., Kodira, C.D., Raffaele, S., Torto-Alalibo, T., Bozkurt, T.O., Ah-Fong, A.M., Alvarado, L., Anderson, V.L., Armstrong, M.R., Avrova, A., Baxter, L., Beynon, J., Boevink, P.C., Bollmann, S.R., Bos, J.I., Bulone, V., Cai, G., Cakir, C., Carrington, J.C., Chawner, M., Conti, L., Costanzo, S., Ewan, R., Fahlgren, N., Fischbach, M.A., Fugelstad, J., Gilroy, E.M., Gnerre, S., Green, P.J., Grenville-Briggs, L.J., Griffith, J., Grunwald, N.J., Horn, K., Horner, N.R., Hu, C.H., Huitema, E., Jeong, D.H., Jones, A.M., Jones, J.D., Jones, R.W., Karlsson, E.K., Kunjeti, S.G., Lamour, K., Liu, Z., Ma, L., Maclean, D., Chibucos, M.C., McDonald, H., McWalters, J., Meijer, H.J., Morgan, W., Morris, P.F., Munro, C.A., O'Neill, K., Ospina-Giraldo, M., Pinzon, A., Pritchard, L., Ramsahoye, B., Ren, Q., Restrepo, S., Roy, S., Sadanandom, A., Savidor, A., Schornack, S., Schwartz, D.C., Schumann, U.D., Schwessinger, B., Seyer, L., Sharpe, T., Silvar, C., Song, J., Studholme, D.J., Sykes, S., Thines, M., van de Vondervoort, P.J., Phuntumart, V., Wawra, S., Weide, R., Win, J., Young, C., Zhou, S., Fry, W., Meyers, B.C., van West, P., Ristaino, J., Govers, F., Birch, P.R., Whisson, S.C., Judelson, H.S., and Nusbaum, C. (2009). Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature **461,** 393-398.

Hall, T.M. (2005). Structure and function of Argonaute proteins. Structure (Camb) **13,** 1403-1408.

Hammond, S.M. (2005). Dicing and slicing The core machinery of the RNA interference pathway. FEBS Lett **579,** 5822-5829.

Havecker, E.R., Wallbridge, L.M., Hardcastle, T.J., Bush, M.S., Kelly, K.A., Dunn, R.M., Schwach, F., Doonan, J.H., and Baulcombe, D.C. (2010). The *Arabidopsis* RNA-directed DNA methylation Argonautes functionally diverge based on their expression and interaction with target loci. Plant Cell **22,** 321-334.

Heisel, S.E., Zhang, Y., Allen, E., Guo, L., Reynolds, T.L., Yang, X., Kovalic, D., and Roberts, J.K. (2008). Characterization of unique small RNA populations from rice grain. PLoS One **3,** e2871.

Henderson, I.R., Zhang, X., Lu, C., Johnson, L., Meyers, B.C., Green, P.J., and Jacobsen, S.E. (2006). Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. Nat Genet **38,** 721-725.

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

Hoffmann, M.H. (2005). Evolution of the realized climatic niche in the genus *Arabidopsis* (Brassicaceae). Evolution **59,** 1425-1436.

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

Huang, X., Wang, J., Aluru, S., Yang, S.P., and Hillier, L. (2003). PCAP: a whole-genome assembly program. Genome Res **13,** 2164-2170.

International Brachypodium Initiative. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature **463,** 763-768.

Jiang, D., Yin, C., Yu, A., Zhou, X., Liang, W., Yuan, Z., Xu, Y., Yu, Q., Wen, T., and Zhang, D. (2006). Duplication and expression analysis of multicopy miRNA gene family members in *Arabidopsis* and rice. Cell Res **16,** 507-518.

John, B., Sander, C., and Marks, D.S. (2006). Prediction of human microRNA targets. Methods Mol Biol **342,** 101-113.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. (2004). Human microRNA targets. PLoS Biol **2,** e363.

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

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res **36,** W5-9.

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

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

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res **110,** 462-467.

Kahveci, T., and Singh, A. (2003). MAP: searching large genome databases. Pac Symp Biocomput **8,** 303-314.

Kasschau, K.D., Xie, Z., Allen, E., Llave, C., Chapman, E.J., Krizan, K.A., and Carrington, J.C. (2003). P1/HC-Pro, a viral suppressor of RNA silencing, interferes with *Arabidopsis* development and miRNA function. Dev Cell **4,** 205-217.

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., and Carrington, J.C. (2007). Genome-wide profiling and analysis of *Arabidopsis* siRNAs. PLoS Biol **5,** e57.

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., Okada, T.N., Siomi, M.C., and Siomi, H. (2008). *Drosophila* endogenous small RNAs bind to Argonaute 2 in somatic cells. Nature **453,** 793-797.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res **12,** 656-664.

Klattenhoff, C., and Theurkauf, W. (2008). Biogenesis and germline functions of piRNAs. Development **135,** 3-9.

Koch, M.A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol Biol Evol **17,** 1483-1498.

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

Kutter, C., Schob, H., Stadler, M., Meins, F., Jr., and Si-Ammour, A. (2007). MicroRNA-mediated regulation of stomatal development in *Arabidopsis*. Plant Cell **19,** 2417-2429.

Lanet, E., Delannoy, E., Sormani, R., Floris, M., Brodersen, P., Crete, P., Voinnet, O., and Robaglia, C. (2009). Biochemical evidence for translational repression by *Arabidopsis* microRNAs. Plant Cell **21,** 1762-1768.

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.

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

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet **11,** 204-220.

Lelandais-Briere, C., Naya, L., Sallet, E., Calenge, F., Frugier, F., Hartmann, C., Gouzy, J., and Crespi, M. (2009). Genome-wide *Medicago truncatula* small RNA analysis revealed novel microRNAs and isoforms differentially regulated in roots and nodules. Plant Cell **21,** 2780-2796.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. Cell **115,** 787-798.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in *Arabidopsis*. Curr Biol **15,** 1501-1507.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: short oligonucleotide alignment program. Bioinformatics **24,** 713-714.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics **25,** 1966-1967.

Liang, H., and Li, W.H. (2009). Lowly expressed human microRNA genes evolve rapidly. Mol Biol Evol **26,** 1195-1198.

Lindow, M., and Krogh, A. (2005). Computational evidence for hundreds of non-conserved plant microRNAs. BMC Genomics **6,** 119.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell **133,** 523-536.

Liu, J., Valencia-Sanchez, M.A., Hannon, G.J., and Parker, R. (2005). MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. Nat Cell Biol **7,** 719-723.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. Science **297,** 2053-2056.

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C., and Green, P.J. (2005a). Elucidation of the small RNA component of the transcriptome. Science **309,** 1567-1569.

Lu, C., Kulkarni, K., Souret, F.F., Muthuvalliappan, R., Tej, S.S., Poethig, R.S., Henderson, I.R., Jacobsen, S.E., Wang, W., Green, P.J., and Meyers, B.C. (2006). MicroRNAs and other small RNAs enriched in the *Arabidopsis* RNA-dependent RNA polymerase-2 mutant. Genome Res **16,** 1276-1288.

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

Lu, J., Shen, Y., Wu, Q., Kumar, S., He, B., Shi, S., Carthew, R.W., Wang, S.M., and Wu, C.I. (2008b). The birth and death of microRNA genes in *Drosophila*. Nat Genet **40,** 351-355.

Lu, S., Sun, Y.H., Shi, R., Clark, C., Li, L., and Chiang, V.L. (2005b). Novel and mechanical stress-responsive microRNAs in *Populus trichocarpa* that are absent from *Arabidopsis*. Plant Cell **17,** 2186-2203.

Ma, Z., Coruh, C., and Axtell, M.J. (2010). *Arabidopsis lyrata* small RNAs: transient *MIRNA* and small interfering RNA loci within the *Arabidopsis* genus. Plant Cell **22,** 1090-1103.

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

Maindonald, J.H., and Braun, J. (2007). Data analysis and graphics using R: an example-based approach. Cambridge University Press, New York, N.Y.

Mallory, A.C., and Vaucheret, H. (2006). Functions of microRNAs and related small RNAs in plants. Nat Genet **38 Suppl,** S31-36.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature **437,** 376-380.

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.

Meister, G., and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. Nature **431,** 343-349.

Mette, M.F., van der Winden, J., Matzke, M., and Matzke, A.J. (2002). Short RNAs can identify new candidate transposable element families in *Arabidopsis*. Plant Physiol. **130,** 6-9.

Meyers, B.C., Axtell, M.J., Bartel, B., Bartel, D.P., Baulcombe, D., Bowman, J.L., Cao, X., Carrington, J.C., Chen, X., Green, P.J., Griffiths-Jones, S., Jacobsen, S.E., Mallory, A.C., Martienssen, R.A., Poethig, R.S., Qi, Y., Vaucheret, H., Voinnet, O., Watanabe, Y., Weigel, D., 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, D.C., Peck, E.A., and Vining, G.G. (2006). Introduction to linear regression analysis. Wiley-Interscience, Hoboken, N.J..

Montgomery, T.A., Howell, M.D., Cuperus, J.T., Li, D., Hansen, J.E., Alexander, A.L., Chapman, E.J., Fahlgren, N., Allen, E., and Carrington, J.C. (2008a). Specificity of ARGONAUTE7-miR390 interaction and dual functionality in *TAS3* trans-acting siRNA formation. Cell **133,** 128-141.

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. (2008b). 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.

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

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

Moxon, S., Jing, R., Szittya, G., Schwach, F., Rusholme Pilcher, R.L., Moulton, V., and Dalmay, T. (2008). Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. Genome Res **18,** 1602-1609.

Nakano, M., Nobuta, K., Vemaraju, K., Tej, S.S., Skogen, J.W., and Meyers, B.C. (2006). Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res **34,** D731-735.

Ning, Z., Cox, A.J., and Mullikin, J.C. (2001). SSAHA: a fast search method for large DNA databases. Genome Res **11,** 1725-1729.

Noma, K., Sugiyama, T., Cam, H., Verdel, A., Zofall, M., Jia, S., Moazed, D., and Grewal, S.I. (2004). RITS acts in cis to promote RNA interference-mediated transcriptional and post-transcriptional silencing. Nat Genet **36,** 1174-1180.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol **302,** 205-217.

Okamura, K., Chung, W.J., Ruby, J.G., Guo, H., Bartel, D.P., and Lai, E.C. (2008). The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. Nature **453,** 803-806.

Ossowski, S., Schneeberger, K., Lucas-Lledo, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science **327,** 92-94.

Oyama, R.K., Clauss, M.J., Formanova´, N., Kroymann, J., Schmid, K.J., Vogel, H., Weniger, K., Windsor, A.J., and Mitchell-Olds, T. (2008). The shrunken genome of *Arabidopsis thaliana*. Plant Syst. Evol. **273,** 257-271.

Pak, J., and Fire, A. (2007). Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. Science **315,** 241-244.

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 USA **102,** 3691-3696.

Park, W., Li, J., Song, R., Messing, J., and Chen, X. (2002). CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. Curr Biol **12,** 1484-1495.

Pearson, W.R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol **183,** 63-98.

Peragine, A., Yoshikawa, M., Wu, G., Albrecht, H.L., and Poethig, R.S. (2004). SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in *Arabidopsis*. Genes Dev **18,** 2368-2379.

Peters, L., and Meister, G. (2007). Argonaute proteins: mediators of RNA silencing. Mol Cell **26,** 611-623.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F.A., van Dyk, L.F., Ho, C.K., Shuman, S., Chien, M., Russo, J.J., Ju, J., Randall, G., Lindenbach, B.D., Rice, C.M., Simon, V., Ho, D.D., Zavolan, M., and Tuschl, T. (2005). Identification of microRNAs of the herpesvirus family. Nat Methods **2,** 269-276.

Piriyapongsa, J., and Jordan, I.K. (2007). A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS ONE **2,** e203.

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

Piriyapongsa, J., Marino-Ramirez, L., and Jordan, I.K. (2007). Origin and evolution of human microRNAs from transposable elements. Genetics **176,** 1323-1337.

Pontier, D., Yahubyan, G., Vega, D., Bulski, A., Saez-Vasquez, J., Hakimi, M.A., Lerbs-Mache, S., Colot, V., and Lagrange, T. (2005). Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in *Arabidopsis*. Genes Dev **19,** 2030-2040.

Qi, Y., He, X., Wang, X.J., Kohany, O., Jurka, J., and Hannon, G.J. (2006). Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. Nature **443,** 1008-1012.

R Development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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.

Rajewsky, N. (2006). microRNA target predictions in animals. Nat Genet **38 Suppl,** S8-13.

Ramachandran, V., and Chen, X. (2008). Degradation of microRNAs by a family of exoribonucleases in *Arabidopsis*. Science **321,** 1490-1492.

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

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.

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

Roger, A.J., and Simpson, A.G. (2009). Evolution: revisiting the root of the eukaryote tree. Curr Biol **19,** R165-167.

Ronemus, M., Vaughn, M.W., and Martienssen, R.A. (2006). MicroRNA-targeted and small interfering RNA-mediated mRNA degradation is regulated by Argonaute, Dicer, and RNA-dependent RNA polymerase in *Arabidopsis*. Plant Cell **18,** 1559-1574.

Rubio-Somoza, I., Cuperus, J.T., Weigel, D., and Carrington, J.C. (2009). Regulation and functional specialization of small RNA-target nodes during plant development. Curr Opin Plant Biol **12,** 622-627.

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell **127,** 1193-1207.

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.

Shabalina, S.A., and Koonin, E.V. (2008). Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol **23,** 578-587.

Sijen, T., Steiner, F.A., Thijssen, K.L., and Plasterk, R.H. (2007). Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. Science **315,** 244-247.

Slotkin, R.K., Freeling, M., and Lisch, D. (2005). Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. Nat Genet **37,** 641-644.

Smalheiser, N.R., and Torvik, V.I. (2005). Mammalian microRNAs derived from genomic repeats. Trends Genet **21,** 322-326.

Song, L., Axtell, M.J., and Fedoroff, N.V. (2010). RNA secondary structural determinants of miRNA precursor processing in *Arabidopsis*. Curr Biol **20,** 37-41.

Song, L., Han, M.H., Lesicka, J., and Fedoroff, N. (2007). *Arabidopsis* primary microRNA processing proteins HYL1 and DCL1 define a nuclear body distinct from the Cajal body. Proc Natl Acad Sci USA **104,** 5437-5442.

Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B **64,** 479-498.

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

Sunkar, R., and Zhu, J.K. (2004). Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. Plant Cell **16,** 2001-2019.

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.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. (2008). The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36,** D1009-1014.

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

Takeda, A., Iwasaki, S., Watanabe, T., Utsumi, M., and Watanabe, Y. (2008). The mechanism selecting the guide strand from small RNA duplexes is different among Argonaute proteins. Plant Cell Physiol **49,** 493-500.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., and Hannon, G.J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature **453,** 534-538.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res. **22,** 4673-4680.

Tibshirani, R. (2006). A simple method for assessing sample sizes in microarray experiments. BMC Bioinformatics **7,** 106.

Tomari, Y., and Zamore, P.D. (2005). Perspective: machines for RNAi. Genes Dev **19,** 517-529.

Tran, R.K., Zilberman, D., de Bustos, C., Ditt, R.F., Henikoff, J.G., Lindroth, A.M., Delrow, J., Boyle, T., Kwong, S., Bryson, T.D., Jacobsen, S.E., and Henikoff, S. (2005). Chromatin and siRNA pathways cooperate to maintain DNA methylation of small transposable elements in *Arabidopsis*. Genome Biol **6,** R90.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA **98,** 5116-5121.

Vagin, V.V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P.D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. Science **313,** 320-324.

Vaucheret, H., Mallory, A.C., and Bartel, D.P. (2006). AGO1 homeostasis entails coexpression of *MIR168* and AGO1 and preferential stabilization of miR168 by AGO1. Mol Cell **22,** 129-136.

Vaucheret, H., Vazquez, F., Crete, P., and Bartel, D.P. (2004). The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. Genes Dev **18,** 1187-1197.

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

Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gasciolli, V., Mallory, A.C., Hilbert, J.L., Bartel, D.P., and Crete, P. (2004). Endogenous trans-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs. Mol Cell **16,** 69-79.

Voinnet, O. (2008). Use, tolerance and avoidance of amplified RNA silencing by plants. Trends Plant Sci **13,** 317-328.

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

Wang, G., and Reinke, V. (2008). A *C. elegans* Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. Curr Biol **18,** 861-867.

Wang, Y., Hindemitt, T., and Mayer, K.F. (2006). Significant sequence similarities in promoters and precursors of *Arabidopsis thaliana* non-conserved microRNAs. Bioinformatics **22,** 2585-2589.

Warburton, P.E., Giordano, J., Cheung, F., Gelfand, Y., and Benson, G. (2004). Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res **14,** 1861-1869.

Warthmann, N., Das, S., Lanz, C., and Weigel, D. (2008). Comparative analysis of the *MIR319a* microRNA locus in *Arabidopsis* and related Brassicaceae. Mol Biol Evol **25,** 892-902.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., Surani, M.A., Sakaki, Y., and Sasaki, H. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature **453,** 539-543.

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.

Wierzbicki, A.T., Haag, J.R., and Pikaard, C.S. (2008). Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell **135,** 635-648.

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

Wright, S.I., Lauga, B., and Charlesworth, D. (2002). Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol Biol Evol **19,** 1407-1420.

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., Wilken, A., and Carrington, J.C. (2005a). DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **102,** 12984-12989.

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

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D., Jacobsen, S.E., and Carrington, J.C. (2004). Genetic and functional diversification of small RNA pathways in plants. PLoS Biol **2,** E104.

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., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R.W., Steward, R., and Chen, X. (2005). Methylation as a crucial step in plant microRNA biogenesis. Science **307,** 932-935.

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.

Zeileis, A., and Hothorn, T. (2002). Diagnostic checking in regression relationships. R News **2,** 7-10.

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

Zhang, X. (2008). The epigenetic landscape of plants. Science **320,** 489-492.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., and Ecker, J.R. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell **126,** 1189-1201.

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

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.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet **39,** 61-69.