The Plant Cell, Vol. 22: 1074–1089, April 2010, www.plantcell.org © 2010 American Society of Plant Biologists

MicroRNA Gene Evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana* [W][OA]

Noah Fahlgren,[a,b] Sanjuro Jogdeo,[a,b] Kristin D. Kasschau,[a,b] Christopher M. Sullivan,[a,b] Elisabeth J. Chapman,[a,b,1] Sascha Laubinger,[c,2] Lisa M. Smith,[c] Mark Dasenko,[a] Scott A. Givan,[a,b] Detlef Weigel,[c] and James C. Carrington[a,b,3]

[a] Center for Genome Research and Biocomputing, Oregon State University, Corvallis, Oregon 97331
[b] Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331
[c] Department of Molecular Biology, Max Planck Institute for Developmental Biology, D-72076 Tübingen, Germany

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*

[1] Current address: Division of Biology, University of California at San Diego, La Jolla, CA 92093.
[2] Current address: Zentrum für Molekularbiologie der Pflanzen, University Tübingen, 72076 Tübingen, Germany.
[3] Address correspondence to carrington@cgrb.oregonstate.edu.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: James C. Carrington (carrington@cgrb.oregonstate.edu).
[W]Online version contains Web-only data.
[OA]Open Access articles can be viewed online without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.110.073999

(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, 2006; 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; see Supplemental Table 1 online). 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 (see Supplemental Figure 2A online). 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* (see Supplemental Figures 1 and 2 online; Lu et al., 2005, 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* (see Supplemental Figure 1 online; (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 (see Supplemental Figure 1 online).

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 (see Supplemental Figures 2B and 2C online). 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; see Supplemental Table 2 online). 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* (see Supplemental Figure 2D online). 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 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 MIRNAs* 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 2). Read data, genomic data, and other information relevant to each *MIRNA* are given in Supplemental Data Set 1A and Supplemental Figure 3 online. 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). Twenty-four *A. lyrata* families had at least two members, whereas 60 families were represented by only one member (Figure 2C). One hundred thirty-four *MIRNA* loci in *A. lyrata* had

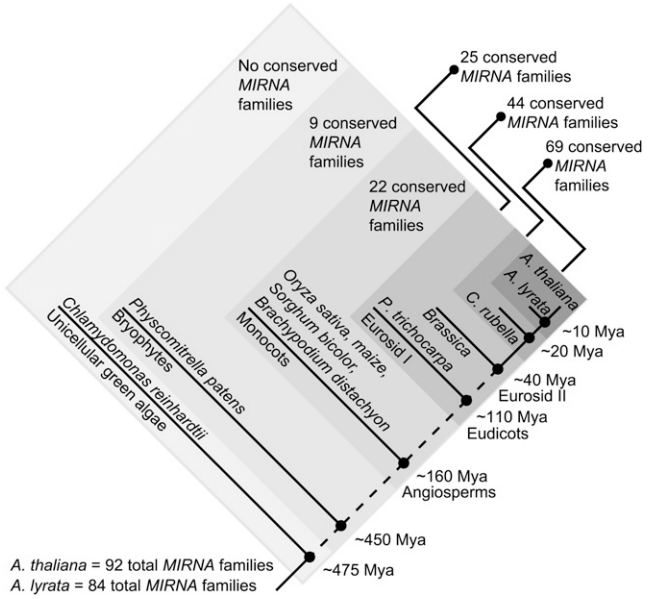

Figure 1. Conservation of *Arabidopsis MIRNA* Families in Plants.

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.

identifiable, conserved orthologous loci in *A. thaliana*, whereas 30 loci were either unique to *A. lyrata* or had diverged (Figures 2A and 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 2A and 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 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 3A and 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 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). 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 3D).

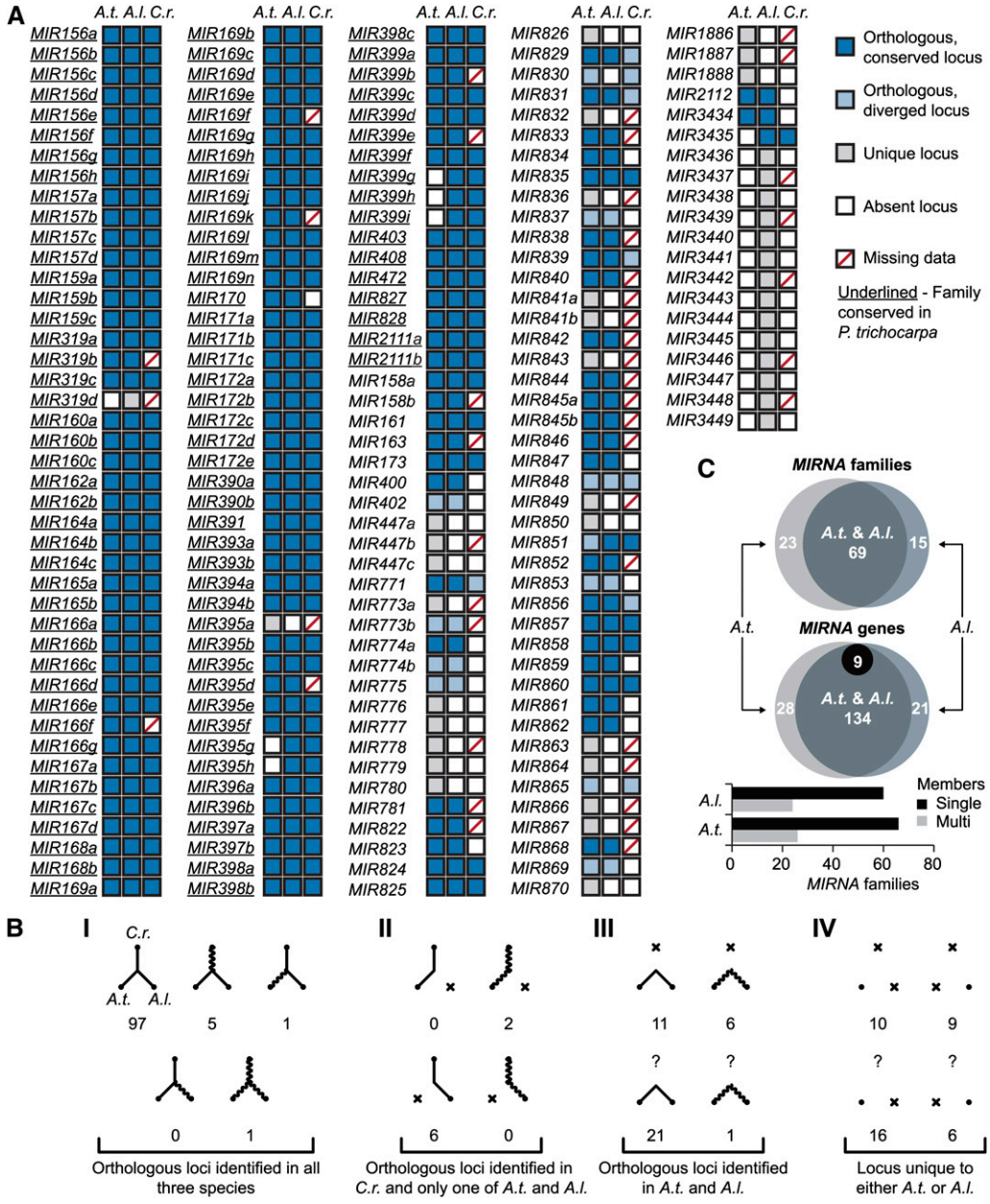

Figure 2. Orthology of *MIRNA* Genes in *A. thaliana*, *A. lyrata*, and *C. rubella*.

(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).

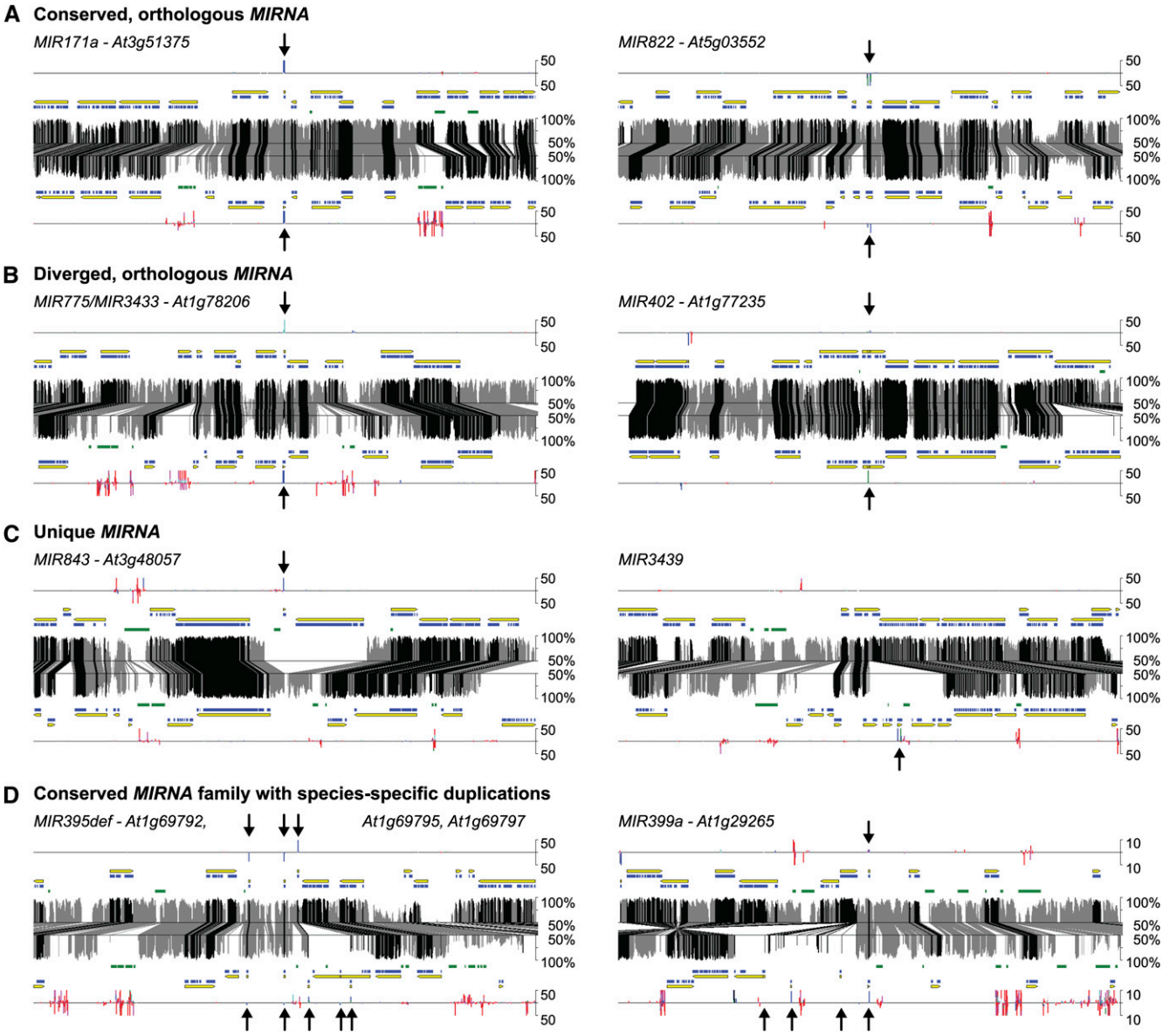

Figure 3. Locus Conservation around Several Classes of *MIRNA* Genes.

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).

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 (see Supplemental Table 1 online). 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). 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 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 2A and 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 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 1 and 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 4A and 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 4A and 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 4A and 4B; see Supplemental Data Set 1C online).

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 4C and 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 4C and 4E). The remaining 21% of *MIRNA* families appeared to be direct duplications of a *MIRNA*-related locus (Figures 4D and 4E). However, in the latter set, each *MIRNA*-related locus contained an ancestral inverted repeat that likely occurred before the *MIRNA*-forming duplication (Figure 4D). In many cases, the duplication extended well beyond the *MIRNA* foldback (Figures 4C and 4D). Regardless of duplication type, nearly one-half of the *MIRNA*-related loci were predicted to be targets of the corresponding miRNA (Figure 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 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 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

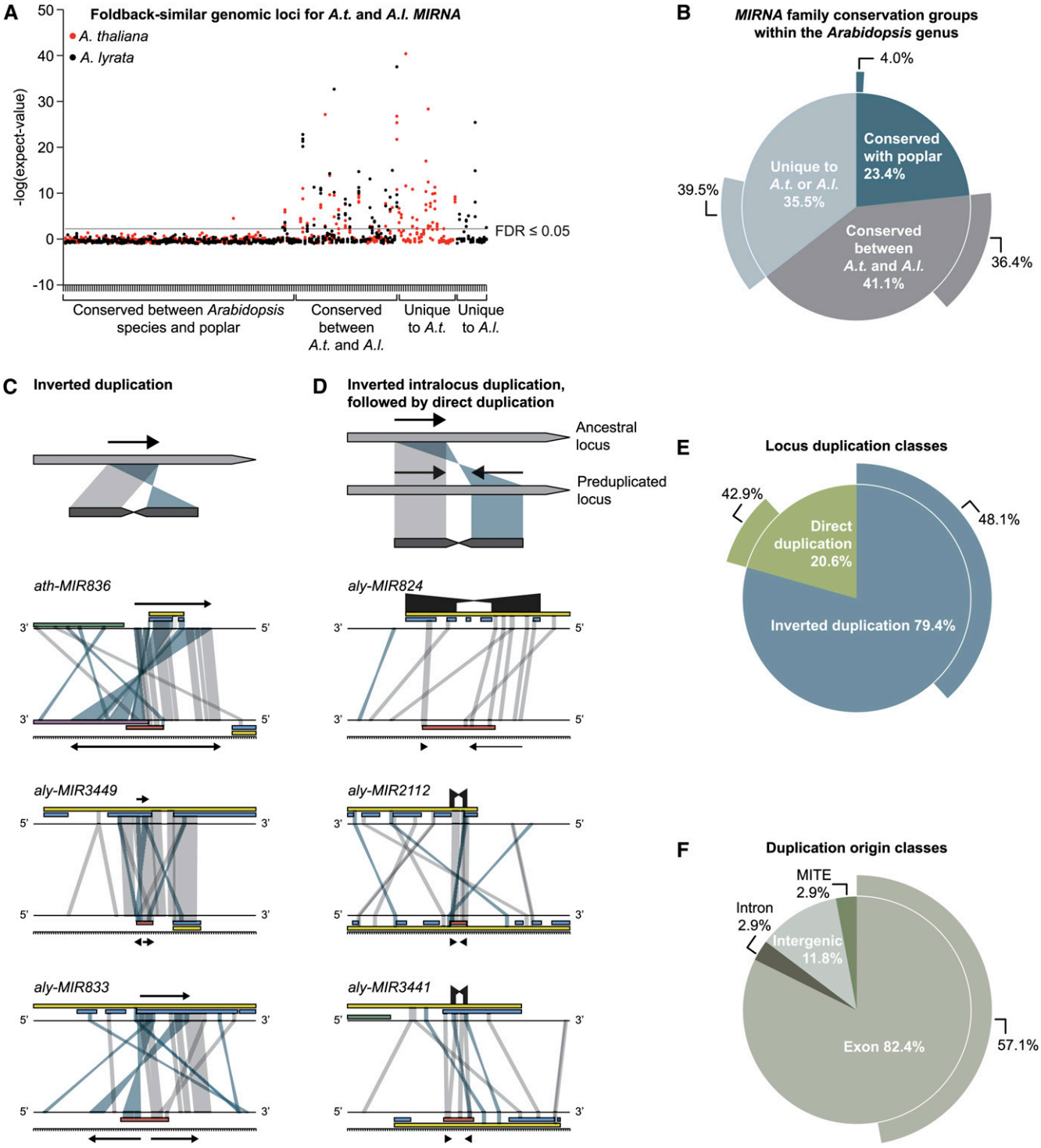

Figure 4. Identification of Intragenomic Loci with Extended Similarity to *MIRNA* Genes.

(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 MIRNAs* aligned to the *A. thaliana* genome are plotted as red dots. *A. lyrata MIRNAs* 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*

(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 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 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 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 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 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 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 6B). In *A. thaliana*, repeat density was more concentrated in the largest intergenic regions (Figure 6B, left), whereas in *A. lyrata*, repeat density was relatively high in average to moderately large intergenic spaces (Figure 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 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 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 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 6D). Repeat density metaplots for most groups

Figure 4. (continued).

and *A. lyrata* but not with *P. trichocarpa*, or *MIRNAs* 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).

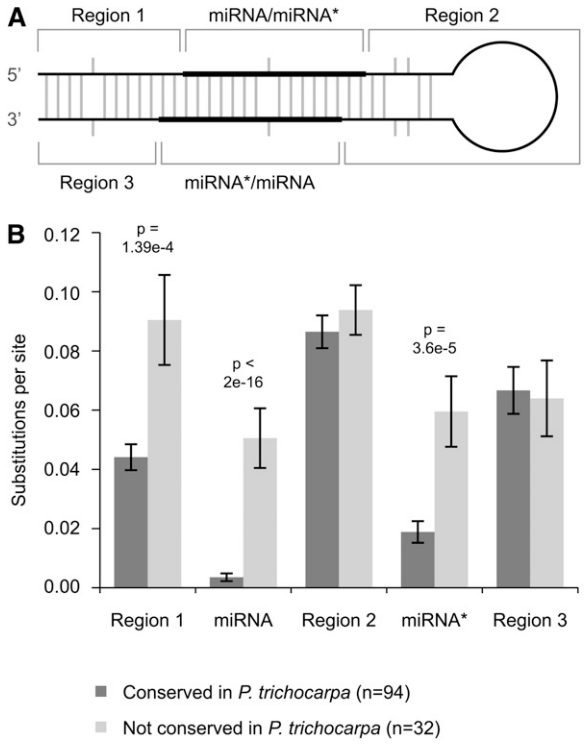

Figure 5. Sequence Divergence within Foldbacks of Orthologous *A. thaliana* and *A. lyrata MIRNA*.

(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.

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 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 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 7A and 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 7A and 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 7A and 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; 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). Orthologs of *A. thaliana* miRNA targets were identified in *A. lyrata* using the MERCATOR/MAVID orthology map. The program TARGET-FINDER (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 8A; see Supplemental Data Set 1D online). Most (146) of the conserved target pairs had target prediction scores of 4 or less in both *A. thaliana* and *A. lyrata* (Figure 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 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 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 8A). In some cases (10),

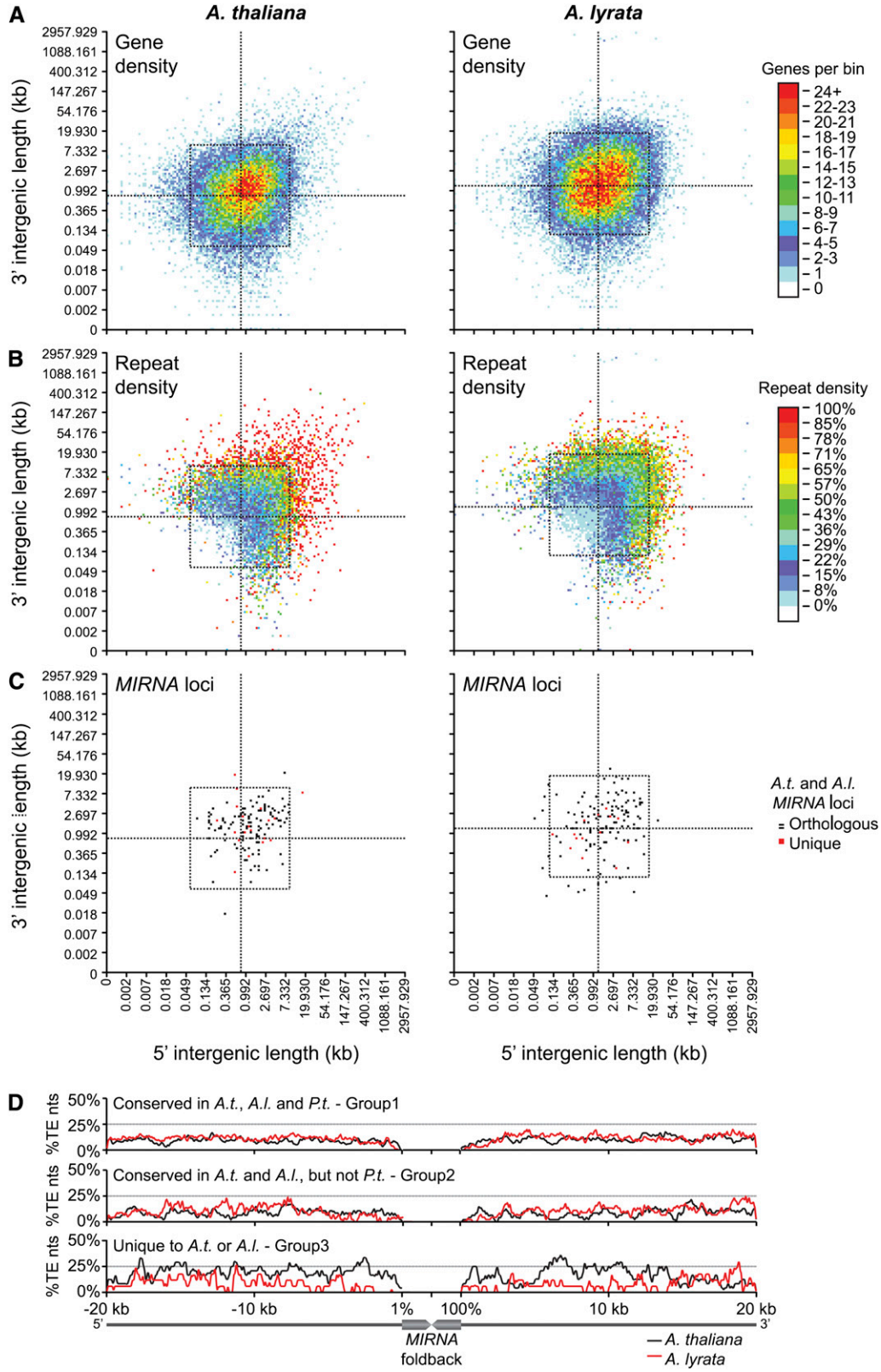

Figure 6. Gene and Repeat Density in the *A. thaliana* and *A. lyrata* Genomes.

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.

the miRNA-target pair was not conserved because the miRNA itself was absent in *A. lyrata* (Figure 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 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 7. For the highly conserved *MIRNA* families (Group 1), 90% of the *A. thaliana* miRNA-target pairs were conserved in *A. lyrata* (Figure 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). 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 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 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 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). miRNA-target pairs involving Group 3 miRNA, which are specific to *A. thaliana*, by definition are all specific to *A. thaliana* (Figure 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, 2008a; Rajagopalan et al., 2006; Axtell et al., 2007; Fahlgren et al., 2007; Molnar et al., 2007; Zhao et al., 2007; Heisel et al., 2008; 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 (see Ma et al., 2010 in this issue). 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

Figure 6. (continued).

(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.

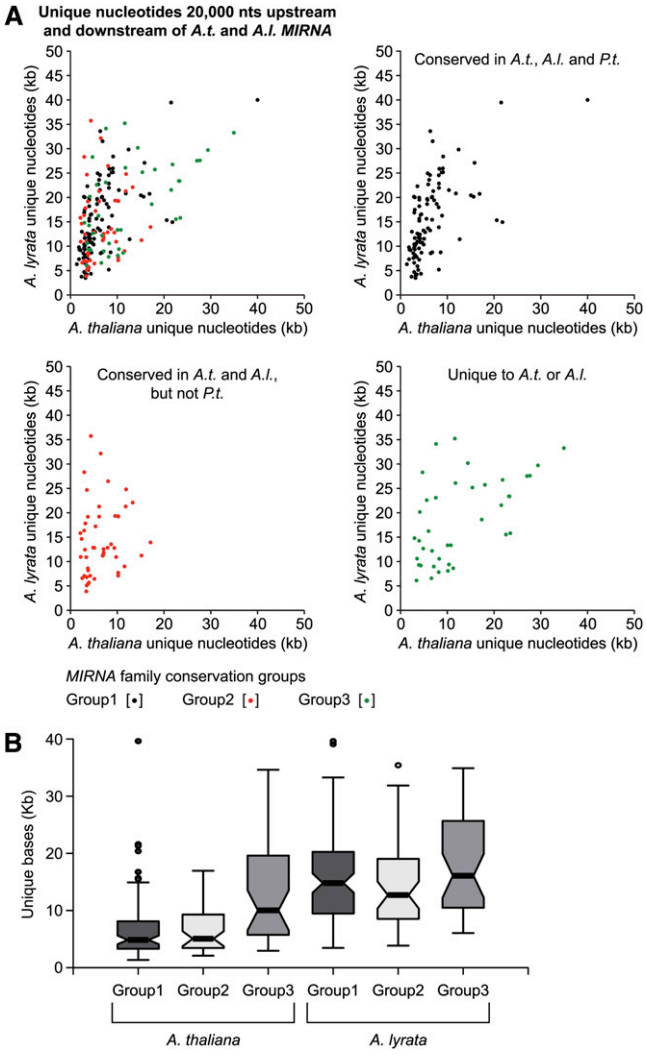

Figure 7. The Genomic Context of *A. thaliana* and *A. lyrata MIRNA*.

(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.

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

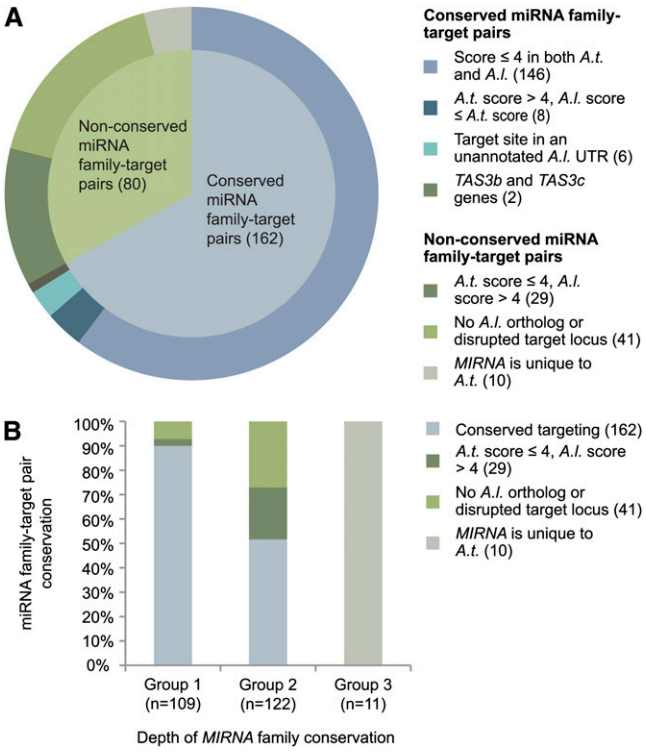

Figure 8. Conservation of *A. thaliana* miRNA-Target Pairs in *A. lyrata.*

(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.

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). 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.

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 Supplemental Figure 1 online. *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.

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.

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 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 Core Development Team, 2009) Q-VALUE package (v1.0; Storey, 2002). For each significant (FDR $\leq$ 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 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 Core Development Team, 2009).

To analyze unique positions flanking *A. thaliana* and *A. lyrata MIRNA* (Figures 7A and 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 Core Development 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 Core Development 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*.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Distribution of Small RNAs, Genes, and Repeat Elements in the *A. lyrata* Genome.

Supplemental Figure 2. Basic Profile of Small RNAs in *A. lyrata* and *A. thaliana*.

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 MIRNAs*.

Supplemental Table 1. Small RNA Library Statistics.

Supplemental Table 2. Summary Statistics for Annotated *A. thaliana* and *A. lyrata* Repeat Features.

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.

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).

Received January 12, 2010; revised March 16, 2010; accepted April 5, 2010; published April 20, 2010.

REFERENCES

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.

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

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.

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

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.

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

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

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

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

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

Clark, R.M., et al. (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.

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.

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

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.

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.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: Tools for microRNA genomics. Nucleic Acids Res. **36:** D154–D158.

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.

Haas, B.J., et al. (2009). Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans.* Nature **461:** 393–398.

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.

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

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

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

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.

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.

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

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.

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.

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.

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

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., et al. (2008a). Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). Proc. Natl. Acad. Sci. USA **105:** 4951–4956.

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

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.

Ma, Z., Coruh, C., and Axtell, M.J. (2010). *Arabidopsis lyrata* small RNAs: Transient *MIRNA* and siRNA loci within the *Arabidopsis* genus. Plant Cell, in press.

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, UK: Cambridge University Press).

Meyers, B.C., et al. (2008). Criteria for annotation of plant microRNAs. Plant Cell **20:** 3186–3190.

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.

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.

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., Formanová, 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.

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

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.

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.

R Core Development Team (2009). R: A Language and Environment for Statistical Computing. (Vienna, Austria: R Foundation for Statistical Computing).

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.

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

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

Smalheiser, N.R., and Torvik, V.I. (2006). Alu elements within human mRNAs are probable microRNA targets. Trends Genet. **22:** 532–536.

Storey, J.D. (2002). A direct approach to false discovery rates. J. R. Stat. Soc., B **64:** 479–498.

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., et al. (2008). The *Arabidopsis* Information Resource (TAIR): Gene structure and function annotation. Nucleic Acids Res. **36:** D1009–D1014.

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.

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. Nucleic Acids Res. **22:** 4673–4680.

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

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.

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.

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.

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.