A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana

Ramya Rajagopalan, Hervé Vaucheret, Jerry Trejo and David P. Bartel

Genes & Dev. 2006 20: 3407-3425
Access the most recent version at doi:10.1101/gad.1476406

A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*

Ramya Rajagopalan,[1] Hervé Vaucheret,[1,2] Jerry Trejo,[1] and David P. Bartel[1,3]

[1]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; [2]Laboratoire de Biologie Cellulaire, Institut Jean-Pierre Bourgin, Institut National de la Recherche Agronomique (INRA), 78026 Versailles Cedex, France

To better understand the diversity of small silencing RNAs expressed in plants, we employed high-throughput pyrosequencing to obtain 887,000 reads corresponding to *Arabidopsis thaliana* small RNAs. They represented 340,000 unique sequences, a substantially greater diversity than previously obtained in any species. Most of the small RNAs had the properties of heterochromatic small interfering RNAs (siRNAs) associated with DNA silencing in that they were preferentially 24 nucleotides long and mapped to intergenic regions. Their density was greatest in the proximal and distal pericentromeric regions, with only a slightly preferential propensity to match repetitive elements. Also present were 38 newly identified microRNAs (miRNAs) and dozens of other plausible candidates. One miRNA mapped within an intron of *DICER-LIKE 1* (*DCL1*), suggesting a second homeostatic autoregulatory mechanism for DCL1 expression; another defined the phase for siRNAs deriving from a newly identified *trans*-acting siRNA gene (*TAS4*); and two depended on DCL4 rather than DCL1 for their accumulation, indicating a second pathway for miRNA biogenesis in plants. More generally, our results revealed the existence of a layer of miRNA-based control beyond that found previously that is evolutionarily much more fluid, employing many newly emergent and diverse miRNAs, each expressed in specialized tissues or at low levels under standard growth conditions.

[*Keywords*: RNA silencing; noncoding RNA; miRNA; siRNA; tasiRNA; high-throughput sequencing]

Supplemental material is available at http://www.genesdev.org.

Received July 31, 2006; revised version accepted November 3, 2006.

Small silencing RNAs direct transcriptional and post-transcriptional gene silencing activities that shape eukaryotic transcriptomes and protein output (Chen 2005; Jones-Rhoades et al. 2006; Mallory and Vaucheret 2006). In plants, these small regulatory RNAs are comprised of microRNAs (miRNAs) and several classes of endogenous small interfering RNAs (siRNAs), which can be differentiated by their distinct modes of biogenesis and the types of genomic loci from which they derive.

The miRNAs derive from primary transcripts that form characteristic stem-loop structures (Ambros 2004; Bartel 2004; Jones-Rhoades et al. 2006). For characterized *Arabidopsis* miRNAs, this miRNA stem-loop precursor is processed by a Dicer-like RNAseIII-type ribonuclease (DCL1) to generate the miRNA:miRNA* duplex, with 2-nucleotide (nt) 3′ overhangs (Park et al. 2002; Reinhart et al. 2002). The miRNA* species derives from the opposing arm of the hairpin and pairs imperfectly to the miRNA (Reinhart et al. 2002). The miRNA strand preferentially incorporates into a silencing complex that has at its core the ARGONAUTE1 (AGO1) protein (Vaucheret et al. 2004; Baumberger and Baulcombe 2005; Qi et al. 2005).

Plant miRNAs have imperfect but extensive complementarity to their mRNA targets, enabling these targets to be predicted with confidence, particularly when the miRNA:target pairing is conserved in multiple species (Rhoades et al. 2002; Jones-Rhoades and Bartel 2004). Plant miRNAs typically direct cleavage of their targets (Llave et al. 2002; Tang et al. 2003). The conserved targets of plant miRNAs are predominantly messages of transcription factors, and the importance of miRNA-mediated regulation of many of these targets for proper embryonic, vegetative, and/or floral development is well established (Chen 2005; Jones-Rhoades et al. 2006; Mallory and Vaucheret 2006). Conserved miRNA targets also include messages for other developmental factors, such as F-box proteins, DCL1, and AGO1, and messages for non-developmental factors, such as stress-response proteins (Chen 2005; Jones-Rhoades et al. 2006; Mallory and Vaucheret 2006).

Endogenous siRNAs derive from long double-stranded RNA (dsRNA) formed as a product of an RNA-dependent RNA polymerase (RdRP), convergent transcription, or transcription of repeats. They typically perform autosilencing, in that they target DNA or transcripts corre-

[3]Corresponding author.
E-MAIL dbartel@wi.mit.edu; FAX (617) 258-6768.
Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.1476406.

sponding to (or related to) the loci from which they derive. Exceptions are the ~21-nt *trans*-acting siRNAs (tasiRNAs), which derive from nonprotein-coding genes known as *TRANS-ACTING siRNA* (*TAS*) genes and post-transcriptionally down-regulate protein-coding transcripts from unrelated loci in a fashion reminiscent of miRNA-directed repression (Peragine et al. 2004; Vazquez et al. 2004; Allen et al. 2005; Yoshikawa et al. 2005). A segment of the *TAS* transcript is converted to dsRNA by RDR6, which is then successively cleaved by DCL4 into 21-nt siRNAs that are then loaded into an AGO1- or AGO7-containing silencing complex where they direct cleavage of the mRNA targets (Peragine et al. 2004; Vazquez et al. 2004; Allen et al. 2005; Gasciolli et al. 2005; Xie et al. 2005b; Yoshikawa et al. 2005; Adenot et al. 2006; Fahlgren et al. 2006; Hunter et al. 2006). One hallmark of tasiRNAs is that they are processed in phase from predominantly one register, which greatly decreases the diversity of siRNAs that accumulate to appreciable levels from a particular TAS locus and ensures production of those with intended targets (Vazquez et al. 2004; Allen et al. 2005). To define the proper phasing register, each of the known *TAS* transcripts is cleaved by a miRNA-programmed silencing complex (Allen et al. 2005; Yoshikawa et al. 2005).

Another type of endogenous siRNA directing PTGS in plants is natural antisense siRNA (nat-siRNA). In the founding example of nat-siRNA-directed silencing, high salt levels induce the expression of *SRO5*, one of a pair of convergently transcribed genes, such that in the presence of transcripts from the other gene, *P5CDH*, a DCL2/RDR6/SGS3/NRPD1a-dependent 24-nt siRNA is produced that directs cleavage of *P5CDH* transcripts (Borsani et al. 2005). This creates a terminus for RdRP production of dsRNA and subsequent processing into secondary siRNAs by DCL1, which can also target *P5CDH* messages (Borsani et al. 2005).

A third type of endogenous siRNA found in plants is heterochromatic siRNA. The concerted activity of plant-specific DNA-dependent RNA polymerases, PolIVa and PolIVb, correlates with the accumulation of 24-nt heterochromatic siRNAs via RDR2-mediated dsRNA formation and DCL3-mediated processing (Xie et al. 2004; Chan et al. 2005). A fraction of these siRNAs associate with AGO4 to form a silencing complex thought to direct sequence-specific methylation events at the DNA and/or chromatin level, which in turn can lead to heterochromatin formation and maintenance at loci from which the siRNAs arise, such as retroelements and the 5S rDNA arrays (Herr et al. 2005; Kanno et al. 2005; Onodera et al. 2005; Pontier et al. 2005). Other siRNAs depend on PolIVa–RDR2–DCL3 but not PolIVb or AGO4, and are not associated with DNA methylation and heterochromatin (Zilberman et al. 2003; Xie et al. 2004; Pontier et al. 2005; Pontes et al. 2006). Their function remains unknown.

Conventional cloning and sequencing of small RNAs from *Arabidopsis* has suggested that plants have a remarkable diversity of endogenous small RNAs (Llave et al. 2002; Park et al. 2002; Reinhart et al. 2002; Sunkar and Zhu 2004; Xie et al. 2004). Recently, massively parallel signature sequencing (MPSS) was employed to obtain a set of 77,434 unique 17-nt signatures of endogenous small RNAs from wild-type *Arabidopsis* (Lu et al. 2005). An appealing alternative that combines the full-length small RNA information of conventional sequencing with the high-throughput character of MPSS is a pyrophosphate-based high-throughput sequencing technique (Margulies et al. 2005). This technique has recently been applied on a pilot scale to the sequencing of *Arabidopsis* small RNAs, generating between 13,000 and 45,000 unique sequences that match the genome (Henderson et al. 2006; Lu et al. 2006; Qi et al. 2006). In order to more broadly characterize the genomic distribution of loci that produce small RNAs in *Arabidopsis*, we used high-throughput pyrosequencing to obtain >340,000 unique small RNA sequences that matched the nuclear, plastid, or mitochondrial genomes. Analysis of this data set provided insights into the evolution, genomics, expression, biogenesis, and function of small silencing RNAs in *Arabidopsis*.

Results

A diverse set of endogenous small RNAs

We adapted our small RNA purification and sequencing protocol, designed to identify RNAs with the size and covalent structure (5′ phosphate and 3′ OH) of DCL products (Lau et al. 2001), to take advantage of high-throughput pyrophosphate sequencing. *Arabidopsis* small RNAs were sequenced from libraries made from whole seedlings, rosette leaves, whole flowers, and siliques. These four runs yielded >1,500,000 reads. Those with recognizable flanking adaptor sequences and with lengths between 16 and 28 nt were compared with *Arabidopsis* nuclear, chloroplast, and mitochondrial genomes. Including another 4239 reads obtained by using conventional methods, 887,266 reads perfectly matched at least one locus and were analyzed further (188,954 from seedling, 186,899 from rosette, 205,649 from flower, and 305,764 from siliques).

These 887,266 reads represented 340,114 unique, although sometimes partially overlapping, sequences (Table 1). About 65% (221,676) of the unique sequences were only sequenced once. The distribution of lengths and 5′ nucleotides for the set of singletons and for the set of sequences with multiple reads were virtually identical (data not shown), suggesting that the two sets represented similar classes of small RNAs. Comparison with a data set of 77,434 unique 17-nt MPSS signatures representing small RNAs that match the *Arabidopsis* genome (Lu et al. 2005) found only 13,596 unique signatures that matched the first 17 nt of at least one of our unique reads. Together, the preponderance of singletons in our library and the modest overlap with the MPSS data set indicated that deep sequencing approaches were still far from saturating the small RNA pools in *Arabidopsis*. Although many small RNAs expressed in *Arabidopsis* remained unidentified, our data set represented a

Table 1. *Summary statistics of small RNAs sequenced from* Arabidopsis

Locus class	Unique sequences		Reads		Mean frequency[a]	Mean hits[b]
Silencing RNAs						
miRBase annotated miRNA hairpin[c]	960	(0.3%)	138,416	(15.6%)	144	1.6
Newly identified miRNA hairpin[c]	361	(0.11%)	7002	(0.8%)	19.4	1.0
Trans-acting siRNA locus	1366	(0.4%)	10,358	(1.2%)	7.6	1.0
Newly identified tasiRNA	34	(0.01%)	111	(0.01%)	3.3	1.0
Candidate silencing RNAs						
rDNA	5318	(1.6%)	20,720	(2.3%)	3.9	24.8
Protein-coding genes						
Sense	18,626	(5.5%)	31,475	(3.5%)	1.7	1.2
Antisense	13,196	(3.9%)	23,441	(2.6%)	1.8	1.2
Sense and antisense	728	(0.2%)	1191	(0.1%)	1.6	1.5
Annotated repeat and mobile elements	111,345	(32.7%)	188,502	(21.2%)	1.7	16.5
Other nuclear genomic	156,407	(46.0%)	293,152	(33.0%)	1.9	1.7
Nonprotein-coding RNAs[d]						
snoRNA	199	(0.06%)	368	(0.04%)	1.8	1.3
snRNA	237	(0.1%)	390	(0.04%)	1.7	5.8
tRNA	2224	(0.7%)	24,606	(2.8%)	11.1	8.5
rRNA	25,639	(7.5%)	134,183	(15.1%)	5.2	3.9
Organellar small RNAs						
Mitochondrial/chloroplast[e]	3474	(1.0%)	13,351	(1.5%)	3.8	n/a
Total	340,114		887,266			

(n/a) Not applicable.
[a]Average number of reads per unique sequence.
[b]Average number of hits to the nuclear genome.
[c]Includes all sequences and reads that mapped to predicted miRNA hairpin precursors.
[d]Matching mature sense strand.
[e]Does not include rRNA or tRNA.

substantial increase in the known diversity of small RNAs matching the genome.

The most abundant reads corresponded to conserved, previously identified miRNAs

As expected, known miRNAs were the sequences most redundantly retrieved from the pool, boasting the highest read frequency of all small RNA classes and 15% of the total (Table 1). All of the miRNA families known to be conserved to poplar and rice (20 families) or just poplar (one additional family; Jones-Rhoades et al. 2006) were represented among our reads, with frequencies as high as 36,093 (miR167). Even the stress-inducible miRNAs miR395 and miR399, previously undetectable in plants grown under normal conditions (Jones-Rhoades and Bartel 2004; Fujii et al. 2005), were represented (13 and 580 reads, respectively), suggesting that some other miRNAs induced in specific conditions might also be represented by multiple reads in our data set. For a few of the miRNAs, including miR319a/b, the sequenced species differed slightly from the annotated species, suggesting refinements of the annotated species (Supplementary Database 1).

Most previously identified conserved miRNA families have multiple, paralogous loci, which combined total 92 loci (Jones-Rhoades et al. 2006). In some cases, members of the same family have slightly different sequences, which can sometimes target distinct sets of messages (Schwab et al. 2005). In other cases, paralogous loci give rise to identical mature miRNAs, raising the question of which paralogs are expressed. Mapping the miRNA* species, rare side products, or degradation fragments unique to a single paralog enabled us to confirm the expression of all but 13 of the 92 loci (Supplementary Database 1), including 19 whose transcription had not been previously confirmed, either by cloning or by mapping the 5′ end of primary transcripts (Xie et al. 2005a). The exceptions were for loci for which no reads could be uniquely mapped.

For most miRNAs, variants of the most abundant read were isolated with 5′ or 3′ heterogeneity, evidenced by missing or extra nucleotides at each terminus (Supplementary Database 1). Occasional slippage of DCL1 processing presumably gives rise to the extra bases, whereas missing nucleotides could result from slippage or end degradation. In contrast to metazoan miRNAs (Lau et al. 2001), we found that 5′ heterogeneity was common in *Arabidopsis* miRNA pools and only slightly less prevalent than 3′ heterogeneity, with no correlation between the extent of heterogeneity and the arm of the foldback that produced the miRNA (Supplementary Database 1). Plant miRNAs might tolerate more extensive 5′ heterogeneity because seed pairing represents a smaller portion of their targeting specificity (Mallory et al. 2004), whereas animal miRNAs truncated or extended by a single nucleotide at their 5′ end would no longer recognize many normal targets and would instead recognize many other messages (Lim et al. 2005).

The miRNA* species had ~9% as many reads as the mature miRNAs, which was higher than the 1% observed in worms (Ruby et al. 2006). This percentage varied widely. For two of the 21 conserved families—miR395 and miR397, represented by 13 and 361 reads, respectively—no miRNA* species were observed (Supplementary Database 1). At the other extreme, miR403* was observed more frequently than mature miR403 (1643 and 66 reads, respectively). Mature miR403 directs cleavage of *AGO2* mRNA (Allen et al. 2005) and is more easily detected by RNA blotting than is miR403* (H. Vaucheret, unpubl.). We infer that sequencing abundance does not always correlate with in vivo abundance, which in any event might not always predict the functional strand.

Several characteristics of *Arabidopsis* miRNAs and their foldbacks emerged from analysis of reads corresponding to miRNA loci that had previously been confidently identified. First, relatively few unique nonoverlapping reads mapped to authentic miRNA foldbacks (Supplementary Database 1). Although DCL1 processing on some foldbacks appeared a little sloppy, it appeared at least globally very precise in that most reads centered on the miRNA/miRNA*, even for stems that were quite extensive. Second, the miRNA* sequence was observed for most loci. Third, <1% of reads mapped to the strand antisense to that giving rise to miRNA and miRNA* (Supplementary Database 1).

Arabidopsis *has many miRNAs lacking close orthologs in other sequenced plants*

In addition to the 21 conserved miRNA families, another five apparently nonconserved miRNA genes (miR158, miR161, miR163, miR173, and miR447) have been confidently identified in *Arabidopsis* (Jones-Rhoades et al. 2006). Each was represented among our reads, with read frequency ranging from 29 for miR447 to 10,573 for miR161 (Supplementary Database 1). How many additional nonconserved miRNAs might exist in *Arabidopsis*? The multitude of other endogenous small RNAs, some of which derive from regions with fortuitous potential to fold into miRNA-like hairpins, has complicated miRNA identification in plants, leading to the suggestion that biogenetic requirements be confirmed using mutant backgrounds prior to annotation (Jones-Rhoades et al. 2006). High-throughput sequencing offered an alternative approach for distinguishing miRNAs from other small RNAs. Candidates from loci with a substantial number of reads deriving from the antisense strand can also be excluded because such antisense reads suggest origin from a perfect dsRNA rather than a hairpin. For the remaining candidates meeting the conventional hairpin-pairing criteria, a sequenced miRNA* species, especially one with 2-nt 3′ overhangs, provides strong evidence that a candidate originates from a DCL-processed stem-loop. As a result, demonstrating that the candidate accumulates in prescribed mutant backgrounds becomes less important, which is particularly helpful for miRNAs difficult to detect on blots.

Using these criteria, we identified 38 additional *Arabidopsis* miRNA families, thereby increasing by 2.5-fold the known diversity of miRNAs in *Arabidopsis* (Fig. 1A,B; Table 2; Supplementary Database 2; Table 2 and Supplementary Database 2 also include a 39th miRNA, miR391, which was absent in miRBase version 7.1). To receive miRNA designation, a miRNA* species (or close variant if the miRNA was sequenced at least three times) must have been observed among the reads, with the exception of miR823, which was validated using RNA blotting and the conventional set of mutants (below). The most abundant read on the foldback was deemed the miRNA, although in cases where read density was roughly equivalent for the most abundant reads from each arm of the foldback, both were together annotated to represent the new miRNA locus (using the 5′ and 3′ designations adopted in similar cases for metazoan miRNAs). Many other miRNAs might exist in *Arabidopsis*; another 40 candidates mapped to plausible hairpins but lacked reads representing the miRNA* species (Supplementary Table 4). Four particularly compelling candidates, each sequenced >25 times (Table 2, CandidateA–CandidateD), were carried forward in subsequent analysis, anticipating that they will eventually be validated.

A search in plant expressed sequence tags (EST) data sets, and the *Oryza sativa* (rice) and *Populus trichocarpa* (poplar) genomes, revealed potential orthologs for only one of the newly identified miRNAs, miR828, which had recognizable orthologs in poplar and leafy spurge, each with one substitution in the mature miRNA (Supplementary Fig. 2). For all other newly identified miRNAs, potential orthologs were either absent in sequenced genomes or found only after relaxing the homology criterion to allow three point substitutions. However, most, if not all, of these candidates appeared to be false positives, because at this stringency an equivalent number of hits was found that satisfied the homology and pairing criteria but mapped to the nonhomologous arms of predicted hairpins. We concluded that most of the newly identified miRNAs do not have identifiable orthologs in the sequence databases and henceforth refer to them all as "nonconserved," while recognizing that a few might have divergent orthologs difficult to identify with confidence, and that many might have orthologs in unsequenced species more closely related to *Arabidopsis thaliana*. Mirroring the search for orthologs, we found no convincing *Arabidopsis* paralogs of the newly identified miRNAs.

Although screening was performed on 20- to 24-nt reads, without preference for a particular length or 5′ nucleotide, 74% of the newly identified miRNA loci encoded a 21-nt miRNA, and 87% encoded a miRNA beginning with a U (Table 2). Thus, these characteristics of the conserved miRNAs (Reinhart et al. 2002) were shared by the newly identified miRNAs. Some intriguing tissue specificities were also evident (Table 2). For example, miR771 and miR839 were sequenced primarily from flowers, miR391 and miR825 appeared preferentially in rosette leaves, miR822 and miR842 were pref-

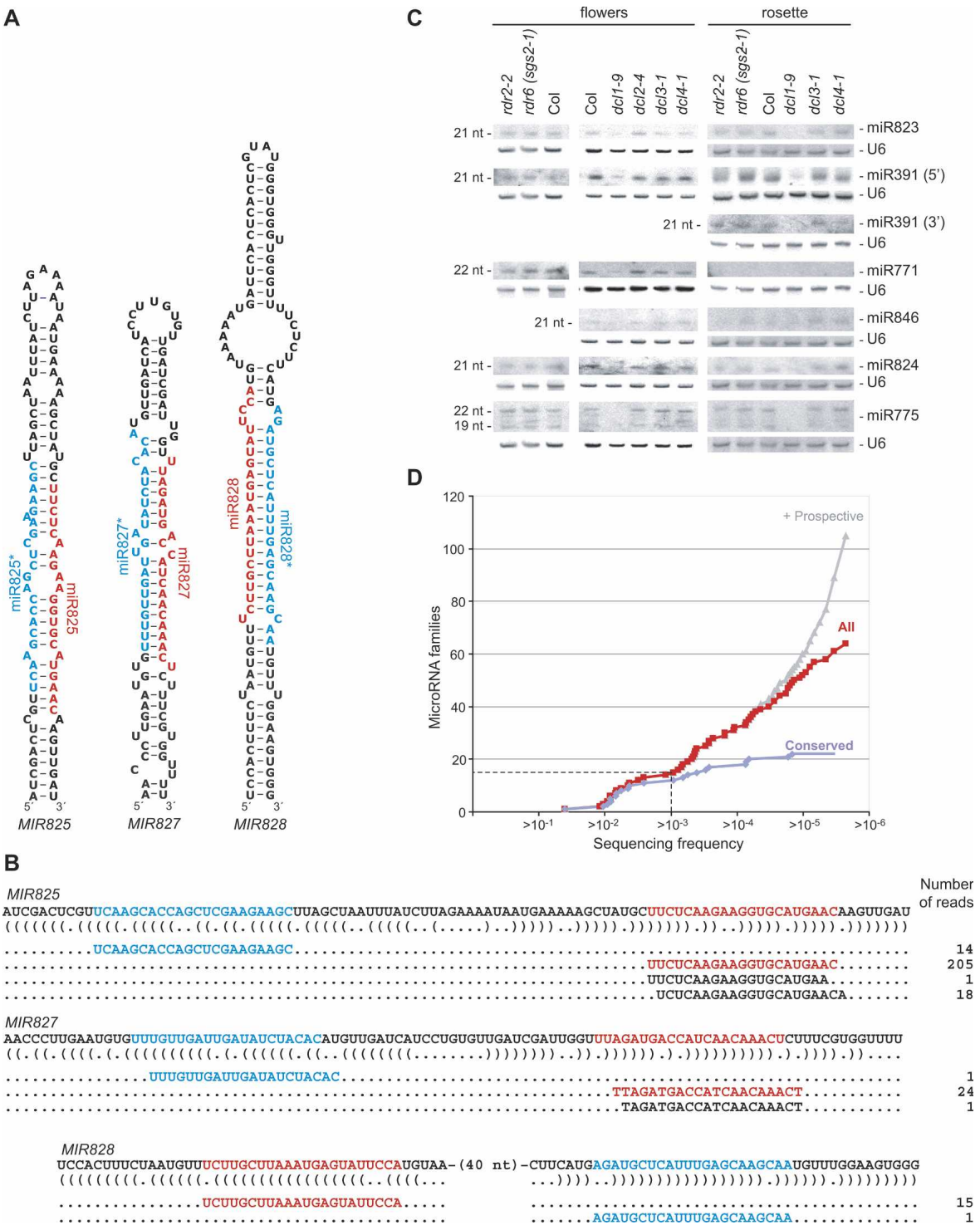

Figure 1. Newly identified *Arabidopsis* miRNAs. (*A*) Predicted secondary structures of miRNA hairpins highlighting the miRNA (red) and miRNA* species (blue). (*B*) The miRNA hairpins of *A*, shown in bracket notation with a tally of reads mapping to the hairpin and nucleotides colored as in *A*. (*C*) RNA blots demonstrating that accumulation of six detectable miRNAs depended on DCL1, not DCL2, DCL3, DCL4, RDR2, or RDR6. As a loading control, blots were stripped and reprobed for U6. (*D*) Sequencing frequencies of *Arabidopsis* miRNA families. Shown are cumulative plots for all *Arabidopsis* miRNA families (red squares), conserved families (violet diamonds), and all families plus the 40 sequenced candidates (gray triangles). Fourteen families, 11 of which were conserved, were sequenced at a frequency of greater than one per 1000 (dashed line).

Table 2. *Newly identified miRNAs in* Arabidopsis

miRNA	Sequence	Len[a]	Se[b]	R[b]	F[b]	Si[b]	Total	miRNA*[c]	Predicted targets[d]	Proteins of targeted messages[e]
miR391(5′)[f]	UUCGCAGGAGAGAUAGCGCCA	21	1	259	12	108	380	Yes		
miR391(3′)	ACGGUAUCUCUCCUACGUAGC	21	13	215	33	53	314	Yes	*At1g72000*	β-fructofuranosidase
miR771[g]	UGAGCCUCUGUGGUAGCCCUCA	22	0	2	109	26	137	Yes		
miR772(5′)[g]	UGUAUGUAUGGUCGAAGUAGG	21	7	1	1	11	20	Yes*	*At2g28010*	Aspartyl protease
miR775[g]	UUCGAUGUCUAGCAGUGCCA	20	97	104	29	155	385	Yes**	*At1g53290*	Galactosyltransferase Avr9 elicitor
miR777[g]	UACGCAUUGAGUUUCGUUGCUU	22	6	2	0	2	10	Yes		
miR779.2[g]	UGAUUGGAAAUUUCGUUGACU	21	17	2	0	1	20	Yes		
miR822	UGCGGGAAGCAUUUGCACAUG	21	922	26	81	34	1063	Yes	*At2g13900, At5g02330*	Four DC1 domain proteins(4)
miR823	UGGGUGGUGAUCAUAUAAGAU	21	305	14	78	296	693	No	*At1g69770*	CMT3 (2)
miR824	UAGACCAUUUGUGAGAAGGGA	21	254	33	46	90	423	Yes	*At3g57230*	AGL16 MADS-box protein
miR825	UUCUCAAGAAGGUGCAUGAAC	21	13	134	1	57	205	Yes		
miR826	UAGUCCGGUUUUGGAUACGUG	21	0	0	16	23	39	Yes	*At4g03060*	AOP2
miR827	UUAGAUGACCAUCAACAAACU	21	0	8	2	14	24	Yes	*At1g02860*	Two SPX C3HC4 RING zinc finger (2)
miR828	UCUUGCUUAAAUGAGUAUUCCA	22	1	0	0	14	15	Yes	*At1g66370, At5g52600*	MYB113, MYB82 (3)
miR829.1	CAAAUUAAAGCUUCAAGGUAG	21	11	0	1	0	12	Yes	*At5g18560*	AP2 domain ethylene response factor
miR829.2	AGCUCUGAUACCAAAUGAUGGAAU	24	10	1	4	0	15	No		
miR830	UAACUAUUUUGAGAAGAAGUG	21	0	1	5	3	9	Yes		
miR831	UGAUCUCUUCGUACUCUUCUUG	22	1	0	0	5	6	Yes	*At3g12190*	Unknown protein
miR832(5′)	UGCUGGGAUCGGGAAUCGAAA	21	0	0	0	6	6	Yes	*At2g46960*	CYP709B1 cytochrome P450
miR832(3′)	UUGAUUCCCAAUCCAAGCAAG	21	0	0	0	3	3	Yes	*At4g30840*	WD-40 protein
miR833(5′)	UGUUUGUUGUACUCGGUCUAGU	22	1	2	0	2	5	Yes		
miR833(3′)	UAGACCGAUGUCAACAAACAAG	22	1	0	2	4	7	Yes		
miR834	UGGUAGCAGUAGCGGUGGUAA	21	0	0	2	1	3	Yes	*At4g00930*	COP1-interacting protein (5)
miR835(5′)	UUCUUGCAUAUGUUCUUUAUC	21	0	0	2	0	2	Yes	*At1g49560*	MYB trancription factor (4)
miR835(3′)	UGGAGAAGAUACGCAAGAAAG	21	1	0	0	1	2	Yes	*At5g46170*	F-box family protein (2)
miR836	UCCUGUGUUUCCUUUGAUGCGUGG	24	0	0	0	2	2	Yes		
miR837(5′)	AUCAGUUUCUUGUUCGUUUCA	21	1	1	0	0	2	Yes	*At1g01160, At4g00850*	Two GIF transcription factors (8)
miR837(3′)	AAACGAACAAAAAACUGAUGG	21	1	0	0	2	3	Yes		
miR838	UUUUCUUCUACUUCUUGCACA	21	0	1	0	1	2	Yes	*At2g45720*	Armadillo/β-catenin protein (6)
miR839	UACCAACCUUUCAUCGUUCCC	21	39	7	184	4	234	Yes*		
miR840	ACACUGAAGGACCUAAACUAAC	22	4	22	3	17	46	Yes*	*At2g02740*	WHIRLY3
miR841	UACGAGCCACUUGAAACUGAA	21	2	13	0	9	24	Yes*	*At2g38810, At4g13570*	Two H2A.F/Z (3)
miR842	UCAUGGUCAGAUCCGUCAUCC	21	14	0	0	0	14	Yes*	*At1g60130, At5g38550*	Five jacalin lectins (7)
miR843	UUUAGGUCGAGCUUCAUUGGA	21	1	4	0	2	7	Yes*	*At3g13830*	Two F-box proteins (3)
miR844(5′)	UGGUAAGAUUGCUUAUAAGCU	21	4	0	0	1	5	Yes*		
miR844(3′)	UUAUAAGCCAUCUUACUAGUU	21	3	2	0	3	8	Yes*	*At3g46540*	Epsin N-terminal homology protein
miR845	CGGCUCUGAUACCAAUUGAUG	21	142	4	153	104	403	Yes**		
miR846	UUGAAUUGAAGUGCUUGAAUU	21	26	51	0	21	98	Yes**	*At2g25980, At5g49850*	Five jacalin lectins (5)
miR847	UCACUCCUCUUCUUCUUGAUG	21	45	2	2	5	54	Yes**	*At1g53720*	Cyclophilin–RNA-interacting protein (6)
miR848	UGACAUGGGACUGCCUAAGCUA	22	16	22	2	11	51	Yes**		
miR849	UAACUAAACAUUGGUGUAGUA	21	12	0	0	4	16	Yes**		
miR850	UAAGAUCCGGACUACAACAAAG	22	1	5	0	9	15	Yes**		
miR851(5′)	UCUCGGUUCGCGAUCCACAAG	21	0	1	9	2	12	Yes**		
miR851(3′)	UGGGUGGCAAACAAAGACGAC	21	0	0	5	4	9	Yes**		
miR852	AAGAUAAGCGCCUUAGUUCUG	21	2	2	0	0	4	Yes**		
miR853	UCCCCUCUUUAGCUUGGAGAAG	22	2	0	0	1	3	Yes**		
CandidateA	UAAUCCUACCAAUAACUUCAGC	22	0	1	52	0	53	No	*At5g41610*	ATCHX18 cation/H+ exchanger (4)
CandidateB	UAGUAACAGAAUUUGGUGUUA	21	0	0	0	40	40	No	*At1g51700*	22 Dof zinc-finger txn factors (22)
CandidateC	UGAGAUGAAAUCUUUGAUUGG	21	8	0	17	6	31	No	*At2g30690*	Unknown protein
CandidateD	UUCGUUGUCUGUUCGACCUUG	21	4	0	16	6	26	No	*At3g08500*	Six MYB transcription factors (7)

[a]Length of mature miRNA sequence.
[b]Number of reads from seedlings (Se), rosette leaves (R), flowers (F), and siliques (Si).
[c]The sequencing of a miRNA* species is denoted "Yes" if the perfect match to the miRNA* (with the 2-nt 3′ overhangs typical of DCL products) was sequenced and was the most abundant read from that arm of the hairpin, "Yes*" if the perfect miRNA* was sequenced but was not the most abundant read from that arm of the hairpin, "Yes**" if only a close heterogeneous variant of the perfect miRNA* was sequenced, and "No" if no star species was recovered.
[d]AGI codes are given for genes with top-scoring target sites for the miRNA. A complete list of predicted mRNA targets of newly identified miRNAs, along with the associated target site alignments and scores, is given in Supplementary Database 4.
[e]Protein products of predicted targets in the best score class. The total number of predicted targets falling within the cutoff is given in parentheses if more than one target was predicted.
[f]Reported as a miRNA in Xie et al. (2005a) but was not annotated in miRBase version 7.
[g]Locus was reported as a miRNA in Lu et al. (2006).

erentially sequenced in seedlings, and miR828 was sequenced most often from siliques. For some with the most striking specificities, we speculate that expression might be at a high level within just a few specialized cells within that organ.

DCL4 processes some Arabidopsis *miRNAs*

Most of the newly identified miRNAs were infrequently recovered by deep sequencing, with median read frequencies of only 13, compared with 731 for the conserved families, suggesting that in plants nonconserved miRNAs are generally expressed at low levels or primarily in specific cells or growth conditions. When RNA from plants grown under normal laboratory conditions was blotted and probed for the 10 most abundant newly identified miRNAs, only eight could be detected using either DNA or LNA probes. Accumulation of six of these eight miRNAs displayed the classical biogenetic profile of DCL1-dependency, with insensitivity to defects in any of the other DCL enzymes or RDR proteins (Fig. 1C).

In contrast to the previously characterized *Arabidopsis* miRNAs, accumulation of two of the eight miRNAs detectable on RNA blots depended on DCL4, not DCL1 (Fig. 2A). One, miR822, was previously classified as an siRNA (ASRP1729) because it accumulates in *dcl1* plants (Allen et al. 2004; Xie et al. 2004). Both miR822 and miR839 were insensitive to defects in RDR2 and RDR6, as expected for RNAs that derive from hairpins rather than dsRNA. Further supporting a hairpin precursor structure was the pattern of reads from these loci (Fig. 2B). Over 99% of reads arose from one strand, with only two of the 1892 *MIR822* reads and one of the 332 *MIR839* reads deriving from the antisense strand—a pattern inconsistent with a perfect dsRNA intermediate. Furthermore, the major species from each arm of the predicted foldbacks paired to each other, with 2-nt 3′ overhangs observed for the miR822:miR822* duplex. Although the cleavage precision did not appear to match that of DCL1, this preferred processing from a localized region of an RNA hairpin stem satisfied the defining feature of miRNAs. We concluded that transcripts from a few miRNA loci are processed by DCL4 rather than by DCL1. The dependency on DCL4 for miR822 and miR839 accumulation appeared even higher than that for tasiRNA accumulation; in the absence of DCL4, tasiRNA precursors are processed into 22-nt and 24-nt species by DCL2 and DCL3, respectively (Gasciolli et al. 2005; Xie et al. 2005b), whereas miR822 and miR839 species are not detectable in either *dcl4-1* or *dcl4-2* plants (Fig. 2A; data not shown).

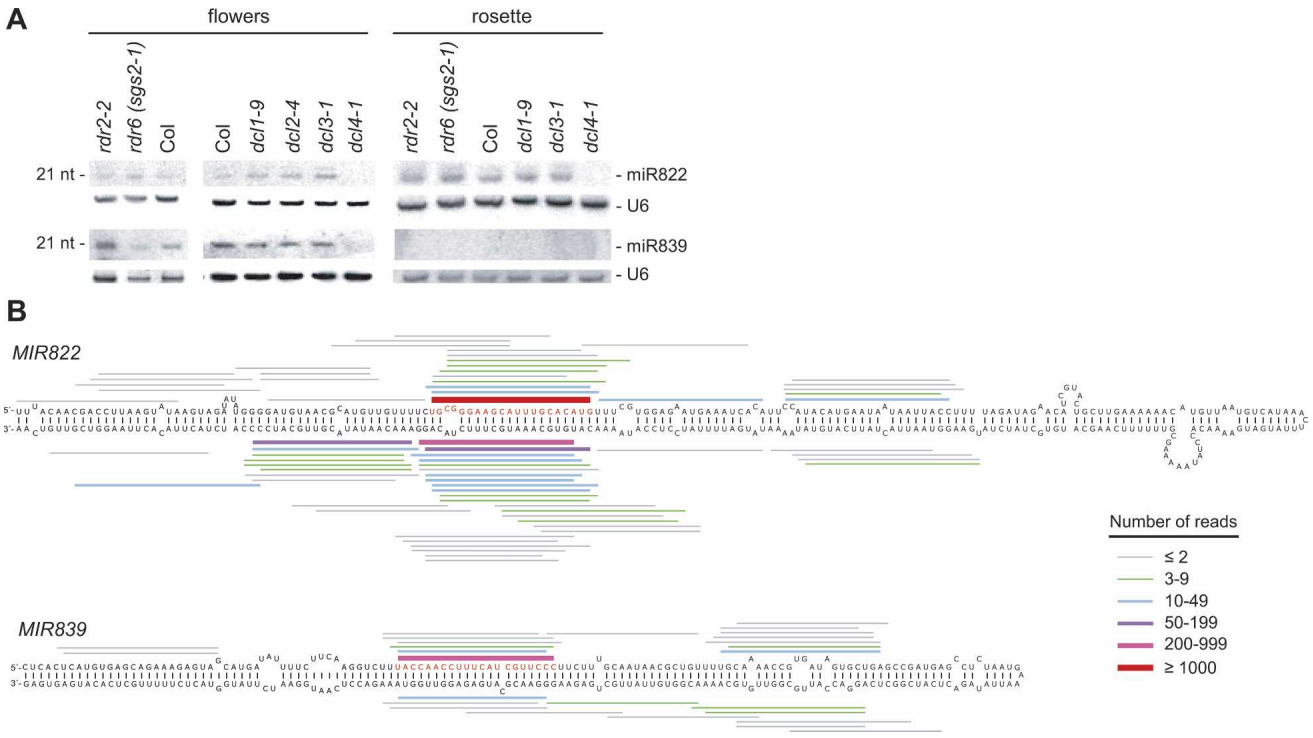

Figure 2. DCL4-dependent miRNAs in *Arabidopsis*. (*A*) RNA blots demonstrating that the accumulation of two detectable miRNAs depended on DCL4 and not on DCL1, DCL2, DCL3, RDR2, or RDR6. As a loading control, blots were stripped and reprobed for U6. Similar results were obtained with *dcl4-1* and *dcl4-2* alleles (data not shown). (*B*) Predicted secondary structures of the miRNA hairpins, with lines denoting the sequences mapping to the miRNA (*top*) and miRNA* (*bottom*) arm of each hairpin. The thickness and color of the lines correspond to the number of total reads representing each small RNA species, as indicated in the key. The two reads corresponding to the antisense of miR822 and the single read mapping to the antisense of miR839 are not depicted.

Predicted targets of newly identified miRNAs

Conserved miRNA targets can be predicted with very high confidence, whereas in single-genome analyses only the more extensively paired interactions can be predicted with reasonable confidence (Jones-Rhoades and Bartel 2004). To better predict nonconserved interactions, scoring rubrics have been developed that preferentially penalize mismatches to the 5′ and central regions of the miRNA (Allen et al. 2005; Schwab et al. 2005), which are more disruptive than those to the 3′ region of the miRNA (Mallory et al. 2004). When applying the rubric of Allen et al. (2005) in a single-genome search to predict targets of 22 unrelated miRNAs, scoring cutoffs that captured 86% of the experimentally confirmed targets of these miRNAs gave a ratio of authentic to false-positive predictions of 6.9:1, estimated by summing the number of targets predicted for the miRNAs and comparing with the average predicted for 10 shuffled cohorts. Using these score cutoffs, we applied the rubric to predict targets of the newly discovered miRNAs, achieving a lower, although still significant, estimated signal:noise ratio of 3.0:1 (Table 2; Supplementary Database 4).

One explanation for the apparently lower specificity was that for six miRNAs, the miRNA and miRNA* species were difficult to distinguish from each other, and thus both were included in the target prediction analysis, recognizing that one of the strands might contribute only false-positive predictions. Similarly, two register-shifted sequences of roughly equal abundance from the miR829 foldback were included. Another explanation might be that some of the newly identified miRNA families have fewer targets with extensive complementarity than do the previously identified families. Indeed, some might not have any biological targets, a subset of which might be "young" DCL1/DCL4 substrates whose processing will soon be lost in the course of neutral evolutionary drift unless a beneficial targeting interaction emerges first. Nonetheless, the prediction of three times as many targets as expected by chance suggested that many of the newly identified miRNAs down-regulate genes. Targets for three of the more abundant miRNAs were validated by 5′ RACE (Supplementary Fig. 3). These were *CMT3*, a miR823 target that encodes a CpNpG DNA cytosine methyltransferase; *AGL16*, a miR824 target that encodes a MADS-box transcription factor; and *MYB113*, a miR828 target that encodes a MYB transcription factor.

Predicted targets of the newly identified miRNAs included transcription factors in the MYB and AP2 families, which each have paralogs known to be targeted by previously identified miRNA families (Supplementary Database 4). In addition, members of transcription factor families not previously known to be regulated by *Arabidopsis* miRNAs, such as MADS-box, ERF (ethylene response factor), WHIRLY, and Dof (DNA binding with one finger) proteins, were among the predicted targets. F-box-containing proteins added to the list of known miRNA targets implicated in protein degradation. A PPR gene distinct from those known to be targeted by miR161, miR400, or *TAS1* or *TAS2 trans*-acting siRNAs was also among the predictions. Other predictions extended the biological processes thought to be regulated by miRNAs. For example, nine jacalin lectins, predicted miR842 and miR846 targets, bind complex carbohydrates and are thought to be involved in initiating pathogen defense responses (Geshi and Brandt 1998). Histone variants, and epigenetic silencing machinery such as CMT3 and a bromo-adjacent homology (BAH) domain-containing protein were predicted targets, suggesting that *Arabidopsis* miRNAs regulate transcriptional silencing pathway components in addition to targeting miRNA biogenetic and effector proteins like DCL1 and AGO1.

Evolutionary origins of miRNA genes

Some miRNAs might have arisen from duplication of their target loci, and if so, those that were recently derived might exhibit similarity to their targets that extends beyond the mature miRNA sequence, as observed previously for miR161 and miR163 (Allen et al. 2004). Six of the newly identified miRNA loci displayed extended sequence similarity with their predicted target genes, diagnostic of common origins. Both arms of the *MIR822* gene were previously observed to have an extended alignment to several DC1 domain-containing genes (Allen et al. 2005). The same pattern was seen for *MIR841* and *MIR826* and their predicted targets (Fig. 3A,B).

A different pattern was observed for *MIR842* and *MIR846*, suggesting an alternative pathway for miRNA gene emergence. As illustrated for *MIR846*, these genes appeared to derive from two regions of their predicted targets, rather than one (Fig. 3C). The simplest explanation for the dual alignment to their targets, with the miRNA arm of the hairpin aligning to one region of the target and the miRNA* arm aligning to the other region, was that a duplication within the targets preceded the duplications that gave rise to the miRNA locus.

Another interesting miRNA–target configuration involved *MIR840*, which was expressed from the opposite strand of its predicted target gene, *AtWhirly3*. This is an arrangement first observed for an Epstein-Barr Virus miRNA and its target (Pfeffer et al. 2004), but one that had not been seen in plants. *AtWhirly3* encodes a homolog of potato p24, a known transcriptional regulator of plant defense and disease resistance genes. In the sense orientation, the miRNA was found within the annotated 3′ untranslated region (UTR) of a PPR mRNA, At2g02750. Although both strands encode a hairpin, our reads did not include any small RNA sequences from the *AtWhirly3* strand. Either the presumptive miRNA or its star sequence could target the *AtWhirly3* 3′ UTR for cleavage. This implies a mechanism by which the expression of one member of a convergent gene pair influences the output of the other—a miRNA counterpart to that observed previously for a convergent gene pair that generates nat-siRNAs (Borsani et al. 2005).

Of the 44 genes for the miRNAs and candidates listed

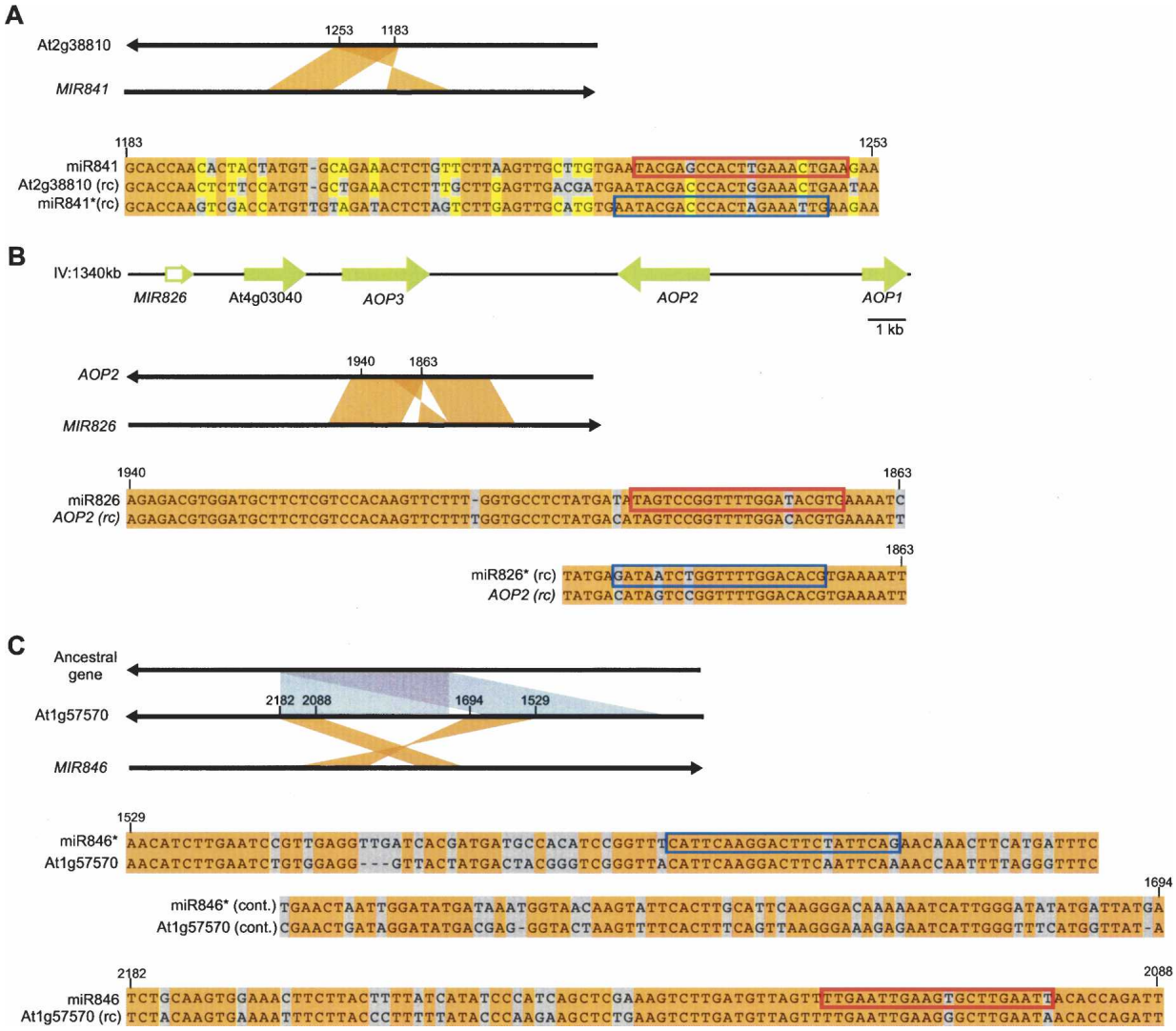

Figure 3. Extended homology between miRNA genes and their predicted target genes, suggestive of common origin. (*A*) *MIR841*, which illustrates a pattern of extended homology (orange shading) resembling that observed previously for *MIR161* and *MIR163* and their respective targets (Allen et al. 2005). Segments corresponding to the mature miRNA (red) and miRNA* (blue) are indicated. The diagram (*top*) depicts the target gene in right–left polarity, and the alignment (*bottom*) depicts the target gene and miRNA* segment as their reverse complements (rc). Numbers indicate positions in the protein-coding gene (At2g38810), counting from its first annotated nucleotide. Nucleotides are shaded to indicate those shared by all (orange) or most (yellow) aligned sequences. (*B*) *MIR826*, for which extended homology suggested an evolutionary pathway whereby a later duplication creating the miRNA* arm was nested within an earlier duplication. The genomic proximity of the miRNA and target gene is shown in the *top* diagram. For the *middle* and *bottom* diagrams, drawing conventions are as in *A*. (*C*) *MIR846*, for which extended homology suggested tandem duplication within an ancestral gene whereby the duplicated regions independently gave rise to the miRNA and miRNA* segments.

in Table 2, 35 were in regions between annotated genes, as is typical of plant miRNA genes (Reinhart et al. 2002), whereas nine overlapped protein-coding genes. One was miR840, described above. Of the remaining eight, miR837, miR838, miR848, miR852, and CandidateD overlapped introns, in the same orientation as the protein-coding host gene—an arrangement that bypasses the need to acquire an independent promoter (Baskerville and Bartel 2005). Mature miR837 also had a second match in the genome, located within the same intron that contains the miR837 stem-loop but in the antisense orientation, suggesting that miR837 might target the pre-mRNA of its host gene, an oligopeptide transporter. miR841 derived from the strand antisense to the intron of At4g13570, a gene closely related to one of its predicted targets but itself not predicted because our search was limited to spliced messages. miR777 and miR834 were localized to the 5′ UTR and 3′ UTR of genes, respectively, with their foldbacks potentially extending into annotated protein-coding regions.

A homeostatic self-regulatory mechanism for DCL1

miR838 derived from a hairpin within intron 14 of the *DCL1* mRNA (Fig. 4A). The foldback potential of this intron was previously noted, and RACE mapping of the *DCL1* transcript revealed a 4.0-kb fragment whose 3′ end terminates at the exon 14/15 junction, and a population of ~2.5-kb fragments, some of which have 5′ ends falling within intron 14 (Xie et al. 2003). Because small RNAs were not detected, the fragments were attributed to aberrant splicing at intron 14 (Xie et al. 2003).

We propose that the presence of this intronic miRNA enables a self-regulatory mechanism that helps maintain *DCL1* homeostasis (Fig. 4B). When nuclear DCL1 protein levels are high, the miRNA biogenesis machinery (including DCL1 and HYL1) could compete more efficiently than the splicing machinery for the *DCL1* precursor transcript. If DCL1 began to process the miRNA hairpin before the intron 14 splice sites were defined and juxtaposed during spliceosome formation, then *DCL1* expression would shift toward a pool of truncated, nonfunctional *DCL1* transcripts, thereby providing a regulatory feedback mechanism that supplements miR162-directed cleavage (Fig. 4C). 5′ RACE confirmed that a population of fragments had 5′ ends terminating at the ends of miR838 (Fig. 4A). The low abundance of the miRNA can be explained by the idea that four linkages must be cut to generate the miRNA:miRNA* duplex, whereas just a single cut bisects the mRNA. Perhaps very little of the duplex is fully excised, and as a result the miRNA never accumulates to sufficient levels to direct efficient target cleavage. We suggest that the processing of other intronic miRNAs might also influence the expression of their host genes—speculation bolstered by the presence of a conserved miRNA-like hairpin in the mammalian *DGCR8* gene, whose protein product functions in pri-miRNA processing (Pedersen et al. 2006).

A newly identified tasiRNA gene in Arabidopsis

In a search for tasiRNA loci, we implemented a clustering algorithm that scanned the genome for phased clusters of ~21-nt reads. This procedure found all five of the previously identified *Arabidopsis* tasiRNA genes (*TAS1a, TAS1b, TAS1c, TAS2*, and *TAS3*) (Supplementary Database 3) and discovered an additional locus, *TAS4* (Fig. 5A), mapping between At3g25800 and At3g25790 (a MYB transcription factor).

Because miRNA-directed cleavage sets the phase for and stimulates production of tasiRNAs (Allen et al.

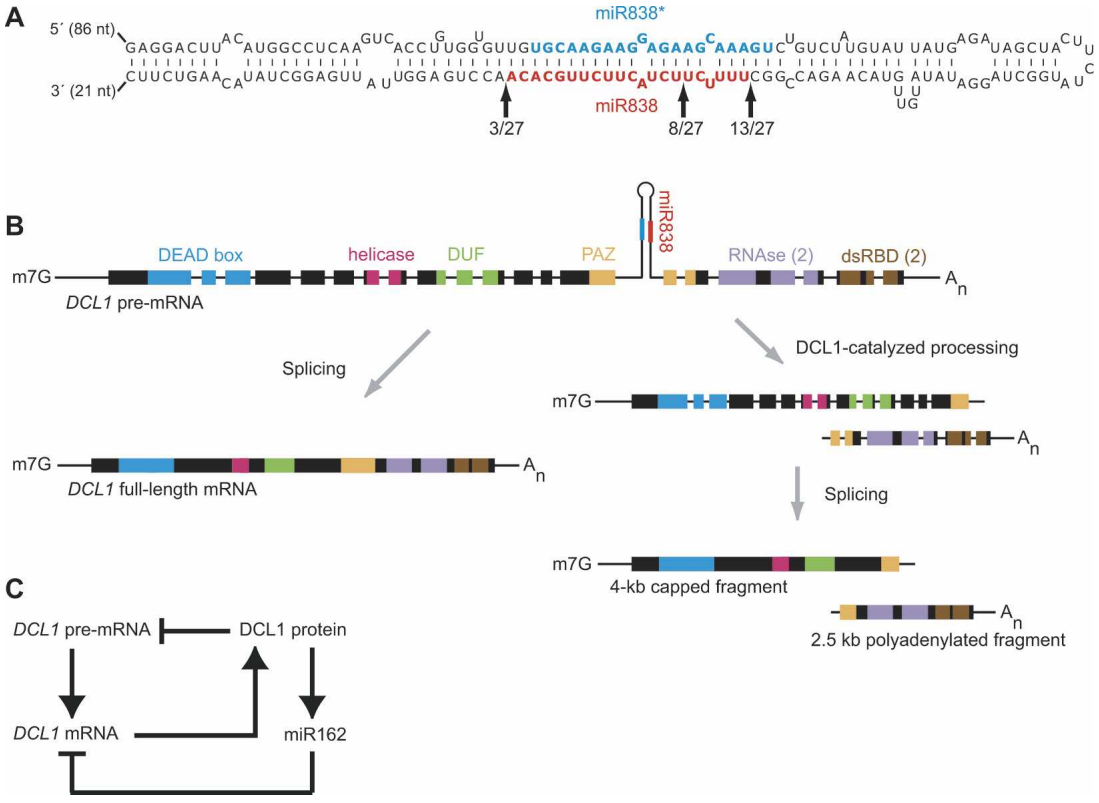

Figure 4. An intronic hairpin positioned so as to mediate DCL1 autoregulation. (*A*) Intron 14 of the *DCL1* primary transcript, and the predicted hairpin structure of miR838. Arrows indicate 5′ ends of RACE-mapped fragments. (*B*) Alternative fates of the *DCL1* primary transcript, which appears to undergo either splicing to generate full-length *DCL1* mRNA or processing by DCL1 itself to generate transcript fragments severed within intron 14. (*C*) A schematic of *DCL1* post-transcriptional autoregulation. When DCL1 protein levels are high, it could compete with splicing machinery for access to intron 14, thereby supplementing miR162-mediated regulation to maintain the proper level of *DCL1* mRNA.

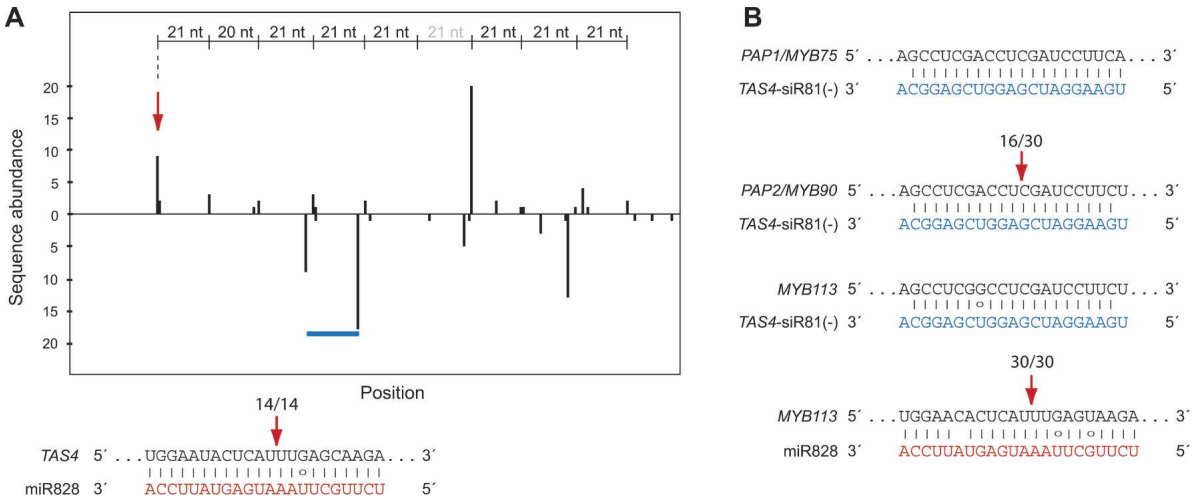

Figure 5. The *TAS4* locus gives rise to tasiRNAs predicted to down-regulate *MYB* transcripts. (*A*) The number of reads with a 5′ terminus at each position is plotted. Bars *above* the axis represent sense reads; those *below* represent antisense reads. The miR828 complementary site is marked by a red arrow and shown *below* the graph, together with the fraction of 5′ RACE clones supporting the indicated cleavage site. *TAS4*-siR81(–), the siRNA predicted to target MYB genes, is indicated (blue bar), as is the spacing separating the phased species at each interval; spacing for the species not represented by reads is indicated in gray. (*B*) *TAS4*-siR81(–) and miR828 complementary sites in three *MYB* genes. Cleavage confirmed by 5′ RACE is indicated (arrows), along with the fraction of clones mapping to the cleavage site. The remaining 14 *PAP2* clones mapped >20 nt from the cleavage site.

2005; Gasciolli et al. 2005; Yoshikawa et al. 2005; Axtell et al. 2006), we searched *TAS4* for miRNA complementary sites upstream of and downstream from the region that generated small RNAs. It identified a single miR828 complementary site. Cleavage at this site, validated by 5′ RACE, defined a 5′ terminus that matched that of the most proximal siRNAs arising from this locus and was in perfect register with the other predominant siRNAs (Fig. 5A). The EST mapping to this region (AU226008) corresponded to the opposite strand of the inferred primary transcript and presumably represented the RDR6-polymerized strand. Although poplar ESTs with miR828 complementary sites were found, conservation of *AtTAS4* to poplar was unclear.

We also predicted three targets for *TAS4*-siR81(–), one of the dominant *TAS4* siRNAs. The predicted targets, *PAP1*/*MYB75*/At1g56650, *PAP2*/*MYB90*/At1g66390, and *MYB113*/At1g66370 (Fig. 5B), encoded three MYB transcription factors that were distinct from the MYB genes targeted by miR159, and those with complementarity to miR835 and CandidateD. PAP1 and PAP2 regulate expression of anthocyanin/flavenoid and phenylpropanoid biosynthetic genes, and might also be involved in regulating leaf senescence (Borevitz et al. 2000; Pourtau et al. 2006). Intriguingly, miR828 was also predicted to down-regulate *MYB113* at an independent target site (Fig. 5B), suggesting a close functional evolutionary relationship among these MYB target genes, miR828, and the *TAS4* cluster. Using 5′ RACE, we identified mRNA cleavage fragments diagnostic of miR828-directed cleavage of *MYB113* and tasiRNA-directed cleavage of *PAP2*, thereby experimentally confirming these predicted targets and demonstrating that the *TAS4* locus was indeed *trans*-acting.

When considered as a group, the 10,469 reads from all six *TAS* genes were predominantly 21 nt and tended to begin with a uridine, as also observed for *Arabidopsis* miRNAs (Supplementary Fig. 1). MicroRNA and synthetic siRNA duplexes assemble into the silencing complex asymmetrically, such that the strand that pairs with less stability at its 5′ terminus is incorporated as the guide strand, while the other strand is degraded (Khvorova et al. 2003; Schwarz et al. 2003). Analysis of the initial 12 siRNA reads from *TAS1a* suggests that tasiRNAs also obey the asymmetry guidelines (Vazquez et al. 2004). The acquisition of many additional tasiRNA reads enabled us to revisit this issue. For each of the six *TAS* genes, each possible duplex represented by a 21mer read (including those out of register with the dominant phasing register) was considered and evaluated for which strand of the duplex yielded more reads. For 57% of the duplexes with energetically distinct terminal base-pairing, the strand with more reads was the one that appeared to be least stably paired at its 5′ terminus. Confounding this analysis, however, was the preference for a U at tasiRNA 5′ termini. If assembly or stability of the silencing complex simply preferred a U at the 5′ terminus of the guide strand, without regard for the differential pairing stabilities at the duplex ends, then there would more frequently be an A:U pair at the 5′ terminus of the guide strand than at the 5′ terminus of the passenger strand, thereby generating artifactual adherence to the pairing asymmetry guidelines. Indeed, the weak adherence vanished when repeating our analysis considering only the duplexes with an A or U at the 5′ termini of both the guide and passenger strands. Apparently, the pairing asymmetry guidelines do not apply for tasiRNAs.

Other endogenous siRNAs in Arabidopsis *mapped predominantly to intergenic regions*

After removing the RNAs that corresponded to the sense strand of ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs) (Table 1; Supplemental Material), and those matching previously annotated and newly identified miRNAs and tasiRNAs, we considered the remaining RNAs that matched the nuclear genome. These included a majority (63%) of the reads and a large majority (90%) of the unique sequences (Table 1). Of the reads that did not match noncoding RNA transcripts, only 10% corresponded solely to annotated mRNAs or introns, which represented substantial depletion when considering that 49% of the sequenced genome is annotated as mRNA and intron. Of those that hit annotated mRNAs or introns, ~46% were exclusively in the antisense orientation to a protein-coding gene. The length and 5′ nucleotide profiles of small RNAs mapping exclusively to the sense strand of genes closely resembled that of small RNAs mapping exclusively to the antisense strand of genes (Supplementary Fig. 1G,H), suggesting that a majority of the sense as well as antisense reads might be siRNAs. We considered them, together with the other small RNAs that did not match noncoding RNA transcripts, as endogenous siRNA candidates.

The candidate siRNAs included 20,720 reads that mapped to the antisense of rRNAs or to ribosomal DNA (rDNA)-like repeats but not to the mature rRNA sequences (Table 1). These are likely to include bona fide siRNAs acting by targeting the rDNA arrays for chromatin or histone modifications (Xie et al. 2004; Pontier et al. 2005; Li et al. 2006; Pontes et al. 2006). Because the fraction of the genome comprised by the rDNA arrays, as well as their copy number, was unknown, and because much of the rDNA sequence was missing from the current assembly, it was difficult to determine if this represented an enrichment.

The candidate siRNAs were mostly 24mers, the size of siRNAs associated with PolIV, heterochromatin formation, and DNA methylation (Chan et al. 2004; Xie et al. 2004; Herr et al. 2005; Kanno et al. 2005; Onodera et al. 2005; Pontier et al. 2005). As expected based on initial sequencing efforts (Tang et al. 2003), the 24mers were enriched for a 5′-terminal adenosine (Supplementary Fig. 1E,F). As for the tasiRNAs, a tendency to adhere to the pairing asymmetry guidelines for silencing complex assembly was observed only in a naïve analysis that did not consider the presumably independent preference for a particular nucleotide at the 5′ termini of the siRNA reads. When correcting for the preference for an A at the 5′ terminus of the candidate siRNAs, the strand with less stable pairing at its 5′ terminus was sequenced no more frequently than expected by chance.

Small RNAs matching annotated protein-coding genes

Some protein-coding genes had a particularly high propensity for spawning small RNAs (Supplementary Table 2). Eleven of the 20 genes most frequently hit, when read counts were normalized by the number of genomic hits and assigned only when unambiguously sense or antisense to genes, were convergently transcribed with a neighboring gene. These included an antisense gene pair, At2g16580/At2g16575, ranked 16th and 19th. Both genes have ORFs with unknown functions, with one ORF falling largely within the intron of the other. Convergent, overlapping transcription presumably generated dsRNA from which the small RNAs were derived. For the nine remaining genes in a convergent context, the 3′ termini were either uncharacterized or had nonoverlapping annotation.

RdRPs provide another mechanism for generating dsRNA. A search for reads that matched a cDNA database but failed to match the genome found 32 reads that spanned mRNA splice junctions in the antisense orientation (Supplementary Table 3). Such reads provided evidence for siRNAs generated by RdRP acting on a spliced mRNA template. The two cDNAs with the most nonoverlapping antisense hits to splice junctions encoded a TIR–NBS–LRR disease-resistance protein (At5g38850) and a basic helix–loop–helix protein (At3g23690). Both were among the top 20 genes hit by small RNAs (Supplementary Table 2). Neither had hits to introns, indicating that for these two genes the RdRP activity acted primarily on spliced templates. However, many of the genes frequently hit by small RNAs had hits to introns. Moreover, many more small RNAs matched the sense strand of mRNA splice junctions (641) than matched the antisense (Supplementary Table 3). The 20-fold difference between sense and antisense reads to splice junctions, compared with the nearly even numbers of sense and antisense reads matching protein-coding genes more generally, suggests that if RdRPs play a major role in producing siRNAs from protein-coding regions, then the templates are usually unspliced transcripts.

Candidate siRNAs derived preferentially from pericentromeric regions, with a slight preference for repeats

When candidate siRNA reads were mapped to each chromosome, plotting sequencing abundance normalized by the number of genome matches (Fig. 6), the largest peak included mostly 22mers and corresponded to a cluster of Gypsy and MuDr elements on the long arm of Chromosome 3. The next two largest peaks corresponded to two rDNA arrays with neighboring repetitive elements on Chromosome 3 and Chromosome 2. Most of the remaining small RNAs were 24mers that mapped to numerous loci dispersed throughout the genome (Fig. 6).

Although the *Arabidopsis* genome is relatively compact, repetitive loci are abundant. They are most dense at and near the centromeres, and their density gradually tapers off in the 2–3 Mb on both sides of each centromere as protein-coding density increases (Fig. 6). Along the remainder of the chromosomes, repeats are present but occur at much lower density. The siRNA density did not

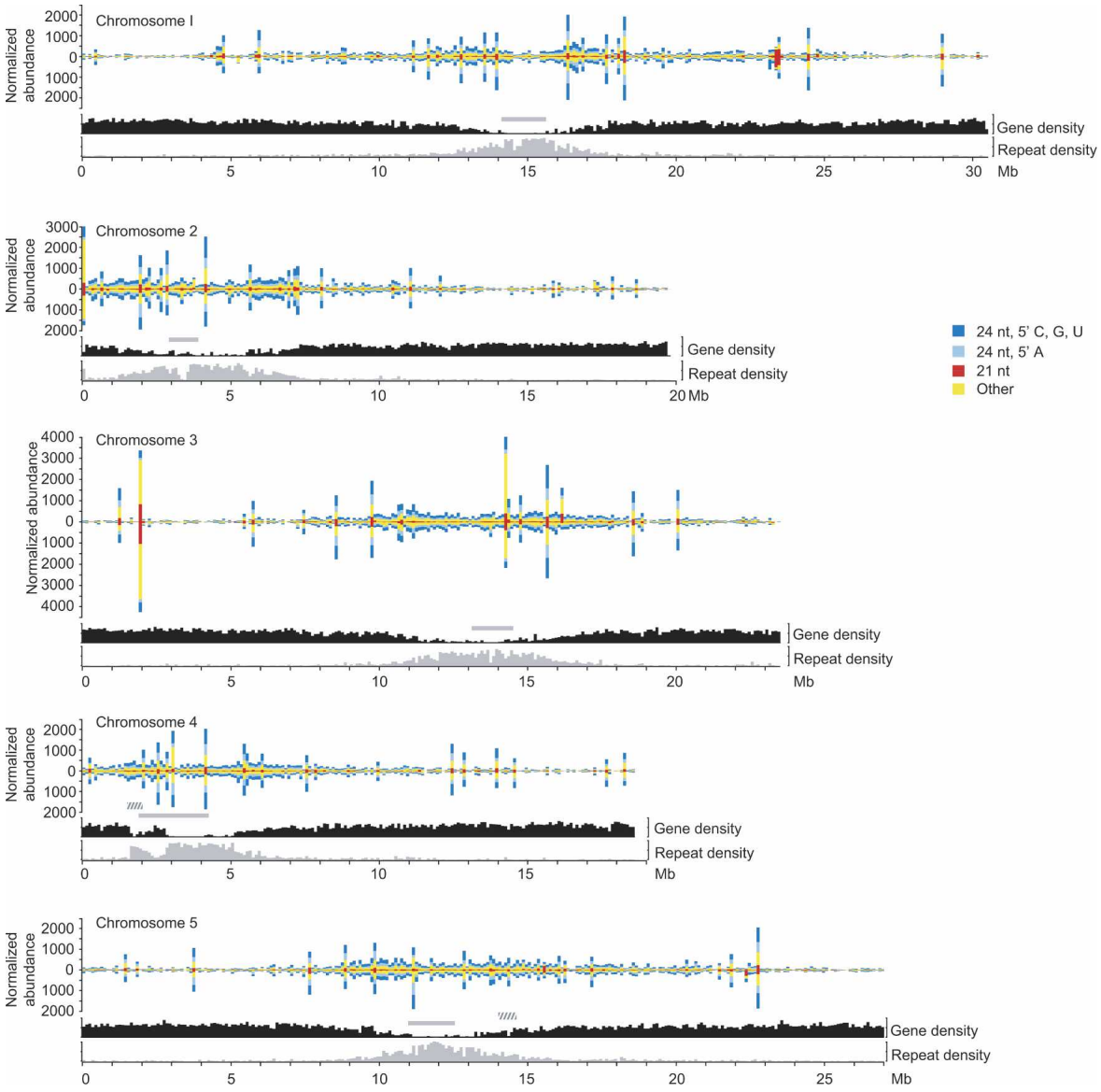

Figure 6. Normalized abundance of candidate siRNAs in 0.1-Mb windows spanning the nuclear genome. Colored bars *above* the axis represent matches to the plus strand; colored bars *below* the axis represent those to the minus strand, with the colors indicating the proportion of 21mers (red), 24mers (light and dark blue), 24mers with a 5′ A (light blue), and other lengths (yellow). *Below* the siRNA profiles are histograms plotting the fraction of nucleotides falling within annotated protein-coding genes (black; scale, 0%–100%) and the fraction falling within repetitive elements annotated by RepeatMasker (gray; scale, 0%–100%). Centromeres are indicated by solid gray bars, and heterochromatic knobs are indicated by hashed gray bars.

peak at the same regions as the repeat density peaked and was instead greatest in the proximal and distal pericentromeric regions, characterized by an intermediate density of both repeats and annotated protein-coding genes (Fig. 6). A look at a diagnostic centromeric repeat class, the ~180-base-pair (bp) repeat satellite arrays (Copenhaver et al. 1999; Nagaki et al. 2003), illustrated this result. Only 2386 reads (1055 unique sequences) matched ~180-bp centromeric satellite repeats annotated by RepeatMasker. This was 0.43% of our reads, whereas the annotated ~180-bp repeat represented 0.39% of the current genome assembly. Because many ~180-bp repeats are missing from the assembly, this slight apparent enrichment was undoubtedly an overestimate; siRNAs deriving from unassembled repeats would artifactually add to the perceived density at any assembled repeats that they match. The unremarkable correspondence between candidate siRNAs and known heterochromatin was also illustrated at the heterochromatic knobs, which were rich in candidate siRNAs, but not more enriched than were the pericentromeric regions that surrounded them (Fig. 6).

The observation that siRNAs were often associated with repeats, but were not highest where repeats were most dense, raised the question of whether siRNAs derived preferentially from repeat loci. Transposons, retro-

elements, and low-complexity sequences identified by RepeatMasker comprised ~15% of the current genome assembly. Of the 558,481 candidate siRNA reads, 188,502 (34%) hit these regions annotated by RepeatMasker—a modest, twofold enrichment over the 15% that would have been expected if the siRNAs derived uniformly from repetitive and nonrepetitive regions throughout the genome. The twofold enrichment was largely attributed to the depletion of both repeats and siRNA matches within annotated protein-coding genes. Of the 51% of the genome that fell between annotated protein-coding genes, ~30% corresponded to repeats annotated by RepeatMasker. Of the 491,180 siRNAs mapping between annotated protein-coding genes, 188,502 (38%) hit regions annotated by RepeatMasker, indicating only a 1.2-fold preference for repeat regions within intergenic regions. The modest preference for repeat regions decreased further when excluding the 20,720 reads deriving from rDNA repeats. When considering only those RNAs associated with AGO4 (Qi et al. 2006), this slight enrichment increased, but not by much. Local (<100 kb) inverted-repeat regions did not appear to be overrepresented among genomic hits of intergenic siRNA candidates that fell outside of repetitive elements identified by RepeatMasker. Of the repetitive DNA detected by RepeatMasker, 80% was either class I (retrotransposon derived) or class II (DNA transposon derived). About 95% of the repeat-associated siRNAs corresponded to these two classes, in the proportion expected based on the contribution of these two classes to the genome. Representation of some of the more well-characterized transposable element families is listed in Supplementary Table 6.

Small RNA hotspots corresponded to unannotated genomic regions

Some intergenic loci had a high propensity to give rise to candidate siRNAs. To supplement the low-resolution analysis (Fig. 6), we performed a higher-resolution search for such siRNA hotspots and then surveyed the annotations corresponding to the top 20, which ranged in length from 0.5 to 50 kb. Although most were in the vicinity of mobile elements or low-complexity sequence, only one hotspot had a transposon at the densest region of siRNAs. Two of the top 20 were very near centromeres, and half were in pericentromeric regions, within 4 Mb of the centromeres. One hotspot, ranked 12th, corresponded to the 5S rDNA array on chromosome 2. Although the topmost-ranked hotspot corresponded to a predicted but unlikely ORF, the other highly ranked hotspots were typically lacking in annotated features within the region producing the majority of small RNAs and represented uncharacterized intergenic regions (Supplementary Fig. 4). A preference for being in local (<100 kb) inverted repeats was not found among the top 20 hotspots, but three lower-ranking loci (ranking 27, 36, and 37) were found in an inverted context. Nine of the top 20 hotspots were in a convergent context with regard to flanking annotated genes. This was higher than might have been expected if convergent, nonconvergent, and divergent contexts were randomly distributed with respect to siRNA-generating loci. However, without mapped transcripts for these convergent flanking genes, the mechanism for siRNA production remains to be elucidated.

Discussion

Endogenous siRNAs in Arabidopsis

Perhaps the most surprising property of the candidate siRNAs was their underwhelming tendency to derive from repeat loci. Of course, RepeatMasker is limited to the identification of repetitive DNA with detectable homology with known repetitive element families, and cannot recognize genomic regions corresponding to novel transposable elements. Although some candidate siRNAs that did not match annotated repeats had multiple genomic matches, suggesting that they might derive from uncharacterized repeats (Table 1), most had only one hit. Moreover, unknown repeats would substantially increase the 1.2-fold enrichment found in intergenic regions only in the unlikely event that the repeats not yet identified were a far richer source of siRNAs than were known repeats.

Part of the reason that repeats generally were not a more rich source of siRNAs was that the regions within and immediately flanking the centromeres, which are mostly annotated repeats, were somewhat depleted in siRNAs when compared with the more distal pericentromeric regions, which had only an intermediate density of repeats (Fig. 6). This observation differed from the report that siRNA density closely mirrors repeat density (Lu et al. 2005). We attribute this apparent contradiction to our normalization of read counts based on the number of times the sequence hit the genome assembly. That is, if a sequence with two reads hit the genome 200 times, we assigned one-hundredth of a count to each locus, rather than two counts to each locus. Our approach attempted to reflect both the fact that a given molecule cannot arise simultaneously from more than one locus, and recent results showing that heterochromatic siRNAs act preferentially at the locus of origin (Buhler et al. 2006), while at the same time leaving ambiguous which repeat locus gave rise to a particular siRNA molecule. Our finding that siRNAs were less abundant at the centromeres, compared with the pericentromeric regions, was reminiscent of the heterochromatic siRNAs of *Schizosaccharomyces pombe*, which map to the heterochromatic outer repeats of the centromeres but not to centromere cores (Reinhart and Bartel 2002). We suggest that heterochromatic siRNAs might function primarily near the boundaries of heterochromatin and euchromatin and play less of a role within large stretches of heterochromatin at the centromeres.

Pairing asymmetry, known to influence incorporation of miRNAs and synthetic siRNAs into silencing complexes (Khvorova et al. 2003; Schwarz et al. 2003), had no detectable correlation with accumulation of tasiRNAs and other *Arabidopsis* siRNA candidates. The same was

found when we analyzed (data not shown) a set of *Arabidopsis* transgene siRNAs previously reported to follow the guidelines (Khvorova et al. 2003). Therefore, for no known cases in animals or plants do endogenously expressed siRNAs preferentially follow the asymmetry guidelines. One explanation might be that most siRNAs in the cell are in the duplex configuration and have not been loaded into the silencing complex. However, no preference was observed when repeating the analysis with a recently reported set of Ago4-associated siRNAs. Therefore, we favor the notion that for most if not all classes of endogenous siRNAs, pairing asymmetry plays little or no role in deciding which strand of the duplex serves as the guide strand. MicroRNAs are a different story; in every plant or animal species examined, miRNA accumulation tends to follow the asymmetry guidelines, even after accounting for their propensity to begin with a U (data not shown). In mammals, synthetic siRNAs also obey the asymmetry guidelines, presumably because vertebrate cells recognize and utilize a synthetic siRNA duplex as if it were an endogenous miRNA duplex. Perhaps more important than pairing asymmetry for endogenous siRNAs is the identity of the 5′ residue. For plant 24mer siRNAs, a 5′-terminal A may favor incorporation or stabilization within the silencing complex, whereas for tasiRNAs, a 5′ U may do the same. The identity of the 5′ nucleotide might also influence the incorporation or stability of miRNAs, which would help explain the strong preferences for U over A and C over G observed at their 5′ termini in all species.

A diverse set of newly emergent miRNAs

Many miRNA candidates have been proposed over the last few years, some of which have been published and annotated in miRBase as authentic *Arabidopsis* miRNAs. Our large data set provided an opportunity to evaluate these candidates and the methods used to identify them. Beyond the 97 confidently identified genes, none of the other current *Arabidopsis* miRNA annotations (miRBase version 7.1) were supported by our data from wild-type plants grown under standard conditions; some of these proposed hairpins matched reads but in a pattern suggestive of endogenous siRNAs (Supplementary Database 1). Furthermore, none of the mature miRNA and candidate sequences of Table 2 matched recently proposed computational candidates, although for seven of 592 recent miRNA predictions (Lindow and Krogh 2005) there was some overlap, which ranged between 7 and 19 nt (Supplementary Database 2). Apart from homologs of known miRNAs, it appears that the only plant miRNAs to have been identified computationally and subsequently confirmed experimentally were those initially reported by Jones-Rhoades and Bartel (2004), at a time when computational searches that required evolutionary conservation could still be productive because some highly conserved miRNAs remained to be found. miR771, miR772, miR775, miR777, miR779, and one of our candidates (CandidateI) corresponded to miRNA hairpins reported while our paper was in review (Lu et al. 2006). For *MIR772*, the species we annotated as the miRNA appears to be the miRNA*; for *MIR779*, the species we sequenced more frequently and annotated as the miRNA derived from a different portion of the hairpin than did miR779.1. Five of our newly identified miRNAs were in a set of 86 candidates previously suggested by analysis of MPSS signatures (Lu et al. 2005) and whose sequences were provided by B. Meyers (pers. comm.).

Of the 38 newly identified miRNAs, only one, miR828, was clearly conserved in other sequenced genomes. The inferred emergence of the new miRNAs after the divergence of the eurosids I (represented by *Arabidopsis*) and II (represented by poplar) ~90 million years ago (Wikstrom et al. 2001) significantly changes our view of miRNAs in plants. Previously, the proportion of known miRNAs that were conserved among eudicots (*Arabidopsis* and poplar) was quite striking—92 of the 97 known genes, 21 of the 26 known families. With respect to the number of microRNA molecules in wild-type plants, this domination by conserved miRNAs still holds, in that >87% of the miRNA molecules we sequenced were conserved throughout sequenced flowering plants. However, with respect to the diversity of plant miRNAs, the picture has dramatically broadened to encompass twice as many nonconserved miRNA families as conserved. In addition to the previously known set of highly conserved miRNAs, each typically expressed from multiple genes at high levels, we now know of a much more evolutionarily flexible set of miRNAs, each expressed from single genes at low levels or in very specialized tissues in plants grown under standard conditions. A plot of the cumulative distribution of sequencing frequency illustrates the relationship between conservation and expression that delineates these two sets of miRNAs (Fig. 1D). All but three of the 14 families sequenced at a frequency of greater than one per 1000 were conserved, whereas only 11 of the 51 families sequenced at a frequency less than one per 1000 appeared to be conserved.

The identification and characterization of these additional miRNAs also expanded our view of plant miRNA biogenesis. At least two miRNAs, miR822 and miR839, depended on DCL4 rather than DCL1 for their accumulation. Just as DCL1, which is primarily responsible for miRNA biogenesis, can generate some siRNAs (Borsani et al. 2005; Bouche et al. 2006; Henderson et al. 2006), DCL4, which is primarily responsible for siRNA biogenesis, can generate some miRNAs. The imprecise cleavage of *MIR161*, which yields miR161 5′ termini ranging over 16 nt, and the dual, apparently sequential processing of the *MIR163* hairpin, which yields miR163.1 and miR163.2 (Kurihara and Watanabe 2004), both illustrate that DCL1 processing of some apparently young miRNA hairpins can be quite heterogeneous (Supplementary Database 1). DCL4-catalyzed cleavage appears even less precise, with a signature yielding numerous minor products often in phase with the miRNA:miRNA* duplex, suggestive of sequential processing after liberation of the miRNA:miRNA* (Fig. 2B). DCL4 can also process per-

fect hairpins to generate transgene siRNAs (Dunoyer et al. 2005), which are presumably far less defined. To the extent that these transgene hairpins might resemble evolutionary precursors of some miRNAs (Allen et al. 2004), we suggest an adaptive switch from DCL4- to DCL1-mediated processing during the course of miRNA gene emergence and evolution, which is driven by selective pressure for enhanced processing precision as the hairpin acquires substitutions and elevated expression, increasing both the probability and consequences of off-target repression. We suspect that the accumulation of some of the miRNAs that accumulate to levels insufficient to detect by RNA blot might also be DCL4 dependent. One attractive candidate would be *MIR841*, which appears to have emerged recently (Fig. 3) and for which register-shifted variants were isolated (Supplementary Database 2).

The nonconserved plant miRNAs presumably emerge and dissipate in short evolutionary time scales. Such rapid emergence of new genes is likely facilitated by the small size and simple architecture of miRNA genes. It could be further facilitated by mechanisms in which they can derive from their future targets (Fig. 3; Allen et al. 2004), although it is unclear whether such mechanisms are relevant for most newly emergent miRNAs or just a minority of them. High-throughput sequencing of small RNAs from species closely related to *Arabidopsis* would help define the life span of these transient miRNA genes as well as the types of processes that they are particularly prone to control. We suspect that these processes will include those under strong positive selection, such as those involved in pathogen response and reproductive isolation.

With the discovery of this diverse, evolutionarily fluid set of miRNAs sequenced at low frequency, the question arises as to how many more miRNAs remain to be reliably identified in *Arabidopsis*. Extrapolating from the sequencing frequencies of the conserved miRNAs, there is little reason to suspect that many more conserved families remain to be discovered (Fig. 1D). Indeed, the curve for the conserved miRNAs was already beginning to plateau with the identification of the first 13 plant miRNA families (Reinhart et al. 2002). The forecast is quite different for the nonconserved families, for which the curve shows no sign of a plateau, particularly when considering the 40 plausible candidates that appeared to derive from miRNA-like hairpins but did not meet our criteria for confident annotation because their miRNA* species had not been sequenced (Fig. 1D, gray symbols; Supplementary Table 1). Based on the large number of genomic segments with predicted potential to give rise to miRNA-like hairpins, it has long been easy to speculate that many nonabundant, nonconserved miRNAs might exist in a given plant or animal. For *Arabidopsis*, such speculation now has experimental support.

Materials and methods

Libraries and sequencing

Wild-type *Arabidopsis* (Columbia accession) plants were grown under standard greenhouse conditions, except seedlings, which were grown as in Reinhart et al. (2002). Total RNA was extracted (Mallory et al. 2001) from whole seedlings, flowers, rosette leaves, and siliques, harvested 6 d, 4 wk, 6 wk, and 2 mo after planting, respectively. Small RNA cDNA libraries were prepared for standard sequencing as in Lau et al. (2001) and for bead-in-well pyrophosphate sequencing as in Axtell et al. (2006). Pyrosequencing was performed at 454 Life Sciences.

Initial processing of reads

cDNA sequences were extracted from raw reads, excluding reads lacking perfect matches to the most proximal 11 nt of both adapter sequences. Unique sequences were mapped to the TAIR/NCBI genome version 6.0 (November 2005), chloroplast and mitochondrial genomes, and rRNA and tRNA sequences from TAIR (http://www.arabidopsis.org), the 5S rRNA Database (http://biobases.ibch.poznan.pl/5SData), the published chromosomal rRNA sequences (Unfried et al. 1989; Unfried and Gruendler 1990), the *Arabidopsis* tRNA database (http://lowelab. ucsc.edu/GtRNAdb/Athal), the *Arabidopsis* snoRNA database (http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/arabidopsis; Brown et al. 2001), miRBase 7.1, and the TIGR v2 *Arabidopsis* Repeat Database. Repetitive regions of the genome were mapped by RepeatMasker (http://www.repeatmasker.org) using AtRepbase (Jurka et al. 2005).

miRNA identification

Twenty-nucleotide to 24-nt sequences with more than one read, 16 or fewer hits to the genome, and no matches to annotated noncoding RNA were folded using RNAfold with 330 nt of upstream and downstream flanking sequence. For efficiency, candidate reads were clustered and only the most abundant in a set of overlapping hits was considered. Structures were evaluated using mirCheck, a script that assesses the quality of a foldback based on a battery of parameters that capture known miRNA hairpins (Jones-Rhoades and Bartel 2004). Hairpins that passed this initial filter were then manually screened. Designation as a miRNA required (1) a foldback in which the duplex region that included 25 nt centered on the most frequently sequenced read had less than eight unpaired nucleotides (summing unpaired nucleotides on both arms of the stem) and no more than three consecutive unpaired nucleotides, of which no more than two were asymmetrically bulged; (2) a sequenced miRNA* species (paired to the miRNA within the duplex with 2-nt 3′ overhangs) or for candidates with three or more reads, a slight variant of the miRNA*; and (3) a sense:antisense read ratio >0.90. In practice, all but six foldbacks that passed manual inspection and were named as miRNA loci had a ratio >99% (Supplementary Database 2). *MIR824*, which has >330 nt between the miRNA and the miRNA*, was found in a separate analysis of genomic regions with abundant 21-nt reads.

Phased siRNA discovery

For each unique small RNA sequence (excluding those matching miRNAs, other noncoding RNAs, or protein-coding genes) a 500-nt window, anchored at one end by that sequence, was evaluated for phased small RNAs. If three or more unique 20- to 23-nt sequences with nonoverlapping hits existed in the window, each was evaluated for phasing with any of the others in the window, allowing ±2 nt of divergence from perfect 21-nt phasing. Phased sequences were extracted from the window and the process was repeated for any remaining 20- to 23-nt sequences until two or fewer unique nonoverlapping hits were left. Each potential phase was then evaluated according to five

parameters: (1) a count of all unique sequences in phase; (2) a count of the reads in the window; (3) a hits-normalized score, whereby the sum of the read frequencies of all phased sequences was divided by the sum of their genomic hits; (4) a normalized 21mer score, which divided the sum of the 21mer reads in phase by the sum of the reads in the window; and (5) a phasing score, which divided the sum of the reads in phase by the sum of the reads in the window. Cutoffs for each score were empirically adjusted to find values that captured all known tasiRNA clusters but restricted the number of false positives.

Target site prediction for miRNAs and TAS4 siRNAs

Patscan was used to search for near matches (up to six mismatches, or four mismatches and one bulged nucleotide) in TAIR version 6.0 *Arabidopsis* cDNA database (http://www.arabidopsis.org) to each miRNA, and target sites were scored as described (Allen et al. 2005). To assess performance, we applied this algorithm to a control set of diverse *Arabidopsis* miRNAs, choosing the most frequently sequenced miRNA variant to represent each known miRNA family for which mRNA targets have been experimentally validated by 5′ RACE, as listed in Jones-Rhoades et al. (2006). We also generated 10 different shuffled cohorts of these 22 miRNAs, preserving dinucleotide composition. Signal:noise ratios were calculated by comparing the total predictions for authentic miRNAs (signal) and the average for shuffled cohorts (noise). Analogously selected cohorts were also used to estimate specificity of target prediction for the newly identified miRNAs.

Hotspot identification

For candidate siRNAs, nonoverlapping 500-bp windows were ranked by scoring small RNA density as a sum of abundances of all sequences in the window, normalized by the sum of the total number of times each sequence hit the genome. Top-ranking windows were then used as seeds for extension in both directions until a 500-bp window lacking any siRNA hits was encountered.

siRNA duplex asymmetry determination

Conceptual duplexes with 2-nt 3′ overhangs were constructed by determining the reverse complement of the sequenced strand. The terminal three pairs (two nearest neighbors) on each end of the duplex were analyzed, comparing sums of the two $-\Delta G^{\circ}_{37}$ nearest-neighbor parameters for RNA duplex stability (Xia et al. 1998). More complex algorithms were also implemented and yielded similar conclusions (Supplemental Material).

RNA gel blot analysis of small RNA expression

RNA gel blots for miRNAs were performed using mutants and protocols as described previously (Vaucheret et al. 2004; Vazquez et al. 2004). Blots were probed with ^{32}P-end-labeled DNA [miR822, miR823, miR391(5′), miR771, miR824, miR775] or LNA [miR391(3′), miR839, miR846] oligonucleotides, each complementary to the entire length of the miRNA.

5′ RACE

5′ RACE was performed as described in Jones-Rhoades and Bartel (2004), except that RNA samples were obtained from whole siliques or seedlings, gene-specific primers (Supplemental Material) were designed to be 70–400 bases from the predicted cleavage site, and the first gene-specific amplifications for *DCL1, PAP2,* and *MYB113* were done with the GeneRacer 5′ outer primer and were followed by two nested amplifications done with the GeneRacer 5′ nested primer.

Accession numbers

All genome-matched small RNA sequences generated in this study are accessible at http://www.ncbi.nlm.nih.gov/geo as Platform GPL3968; Samples GSM118372, GSM118373, GSM118374, and GSM118375; and Series GSE5228. All genomic loci of the small RNAs are listed in Supplementary Table 5; sequences that hit cDNA but not the genome are listed in Supplementary Table 3.

Acknowledgments

We thank M. Axtell for many helpful discussions, G. Bell for computational advice, A. Hall and J. Jurek (Preuss Laboratory) for pilot analysis of centromeric candidate siRNAs, and M. Jones-Rhoades and G. Ruby for scripts. This work was supported by the Prix Louis D. award from the Institut de France and a grant from the NIH to D.B.

References

Adenot, X., Elmayan, T., Lauressergues, D., Boutet, S., Bouche, N., Gasciolli, V., and Vaucheret, H. 2006. DRB4-dependent *TAS3 trans*-acting siRNAs control leaf morphology through AGO7. *Curr. Biol.* **16:** 927–932.

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.

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.

Ambros, V. 2004. The functions of animal microRNAs. *Nature* **431:** 350–355.

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

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

Baskerville, S. and Bartel, D.P. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11:** 241–247.

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.* **102:** 11928–11933.

Borevitz, J.O., Xia, Y., Blount, J., Dixon, R.A., and Lamb, C. 2000. Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell* **12:** 2383–2394.

Borsani, O., Zhu, J., Verslues, P.E., Sunkar, R., and Zhu, J.K. 2005. Endogenous siRNAs derived from a pair of natural *cis*-antisense transcripts regulate salt tolerance in *Arabidopsis. Cell* **123:** 1279–1291.

Bouche, N., Lauressergues, D., Gasciolli, V., and Vaucheret, H. 2006. An antagonistic function for *Arabidopsis* DCL2 in development and a new function for DCL4 in generating viral siRNAs. *EMBO J.* **25:** 3347–3356.

Brown, J.W., Clark, G.P., Leader, D.J., Simpson, C.G., and Lowe, T. 2001. Multiple snoRNA gene clusters from *Arabidopsis*.

RNA **7:** 1817–1832.
Buhler, M., Verdel, A., and Moazed, D. 2006. Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing. *Cell* **125:** 873–886.
Chan, S.W., Zilberman, D., Xie, Z., Johansen, L.K., Carrington, J.C., and Jacobsen, S.E. 2004. RNA silencing genes control de novo DNA methylation. *Science* **303:** 1336.
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.
Chen, X. 2005. MicroRNA biogenesis and function in plants. *FEBS Lett.* **579:** 5923–5931.
Copenhaver, G.P., Nickel, K., Kuromori, T., Benito, M.I., Kaul, S., Lin, X., Bevan, M., Murphy, G., Harris, B., Parnell, L.D., et al. 1999. Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286:** 2468–2474.
Dunoyer, P., Himber, C., and Voinnet, O. 2005. DICER-LIKE 4 is required for RNA interference and produces the 21-nucleotide small interfering RNA component of the plant cell-to-cell silencing signal. *Nat. Genet.* **37:** 1356–1360.
Fahlgren, N., Montgomery, T.A., Howell, M.D., Allen, E., Dvorak, S.K., Alexander, A.L., and Carrington, J.C. 2006. Regulation of AUXIN RESPONSE FACTOR3 by *TAS3* ta-siRNA affects developmental timing and patterning in *Arabidopsis*. *Curr. Biol.* **16:** 939–944.
Fujii, H., Chiou, T.J., Lin, S.I., Aung, K., and Zhu, J.K. 2005. A miRNA involved in phosphate-starvation response in *Arabidopsis*. *Curr. Biol.* **15:** 2038–2043.
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.
Geshi, N. and Brandt, A. 1998. Two jasmonate-inducible myrosinase-binding proteins from *Brassica napus* L. seedlings with homology to jacalin. *Planta* **204:** 295–304.
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.
Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. 2005. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308:** 118–120.
Hunter, C., Willmann, M.R., Wu, G., Yoshikawa, M., de la Luz Gutierrez-Nava, M., and Poethig, S.R. 2006. *Trans*-acting siRNA-mediated repression of ETTIN and ARF4 regulates heteroblasty in *Arabidopsis*. *Development* **133:** 2973–2981.
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.
Kanno, T., Huettel, B., Mette, M.F., Aufsatz, W., Jaligot, E., Daxinger, L., Kreil, D.P., Matzke, M., and Matzke, A.J. 2005. Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nat. Genet.* **37:** 761–765.
Khvorova, A., Reynolds, A., and Jayasena, S.D. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115:** 209–216.
Kurihara, Y. and Watanabe, Y. 2004. *Arabidopsis* micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc. Natl. Acad. Sci.* **101:** 12753–12758.
Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858–862.
Li, C.F., Pontes, O., El-Shami, M., Henderson, I.R., Bernatavichute, Y.V., Chan, S.W., Lagrange, T., Pikaard, C.S., and Jacobsen, S.E. 2006. An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in *Arabidopsis thaliana*. *Cell* **126:** 93–106.
Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769–773.
Lindow, M. and Krogh, A. 2005. Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics* **6:** 119.
Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14:** 1605–1619.
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, 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., et al. 2006. MicroRNAs and other small RNAs enriched in the *Arabidopsis* RNA-dependent RNA polymerase-2 mutant. *Genome Res.* **16:** 1276–1288.
Mallory, A.C. and Vaucheret, H. 2006. Functions of microRNAs and related small RNAs in plants. *Nat. Genet.* **38 (Suppl. 1):** S31–S36.
Mallory, A.C., Ely, L., Smith, T.H., Marathe, R., Anandalakshmi, R., Fagard, M., Vaucheret, H., Pruss, G., Bowman, L., and Vance, V.B. 2001. HC-Pro suppression of transgene silencing eliminates the small RNAs but not transgene methylation or the mobile signal. *Plant Cell* **13:** 571–583.
Mallory, A.C., Reinhart, B.J., Jones-Rhoades, M.W., Tang, G., Zamore, P.D., Barton, M.K., and Bartel, D.P. 2004. MicroRNA control of *PHABULOSA* in leaf development: Importance of pairing to the microRNA 5′ region. *EMBO J.* **23:** 3356–3364.
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., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.
Nagaki, K., Talbert, P.B., Zhong, C.X., Dawe, R.K., Henikoff, S., and Jiang, J. 2003. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics* **163:** 1221–1225.
Onodera, Y., Haag, J.R., Ream, T., Nunes, P.C., Pontes, O., and Pikaard, C.S. 2005. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120:** 613–622.
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.
Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., and Haussler, D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2:** e33.
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.

Pfeffer, S., Zavolan, M., Grasser, F.A., Chien, M., Russo, J.J., Ju, J., John, B., Enright, A.J., Marks, D., Sander, C., et al. 2004. Identification of virus-encoded microRNAs. *Science* **304:** 734–736.

Pontes, O., Li, C.F., Nunes, P.C., Haag, J., Ream, T., Vitins, A., Jacobsen, S.E., and Pikaard, C.S. 2006. The *Arabidopsis* chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* **126:** 79–92.

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.

Pourtau, N., Jennings, R., Pelzer, E., Pallas, J., and Wingler, A. 2006. Effect of sugar-induced senescence on gene expression and implications for the regulation of senescence in *Arabidopsis*. *Planta* **224:** 556–568.

Qi, Y., Denli, A.M., and Hannon, G.J. 2005. Biochemical specialization within *Arabidopsis* RNA silencing pathways. *Mol. Cell* **19:** 421–428.

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.

Reinhart, B.J. and Bartel, D.P. 2002. Small RNAs correspond to centromere heterochromatic repeats. *Science* **297:** 1831.

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.

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

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

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

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

Tang, G., Reinhart, B.J., Bartel, D.P., and Zamore, P.D. 2003. A biochemical framework for RNA silencing in plants. *Genes & Dev.* **17:** 49–63.

Unfried, I. and Gruendler, P. 1990. Nucleotide sequence of the 5.8S and 25S rRNA genes and of the internal transcribed spacers from *Arabidopsis thaliana*. *Nucleic Acids Res.* **18:** 4011.

Unfried, I., Stocker, U., and Gruendler, P. 1989. Nucleotide sequence of the 18S rRNA gene from *Arabidopsis thaliana* Co10. *Nucleic Acids Res.* **17:** 7513.

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., 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.

Wikstrom, N., Savolainen, V., and Chase, M.W. 2001. Evolution of the angiosperms: Calibrating the family tree. *Proc. Biol. Sci.* **268:** 2211–2220.

Xia, T., SantaLucia Jr., J., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., and Turner, D.H. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37:** 14719–14735.

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., 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.

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

Xie, Z., Allen, E., Wilken, A., and Carrington, J.C. 2005b. DICER-LIKE 4 functions in *trans*-acting small interfering RNA biogenesis and vegetative phase change in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* **102:** 12984–12989.

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.

Zilberman, D., Cao, X., and Jacobsen, S.E. 2003. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299:** 716–719.