

Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*

Edwards Allen^{1,2}, Zhixin Xie^{1,2}, Adam M Gustafson^{1,2}, Gi-Ho Sung², Joseph W Spatafora² & James C Carrington^{1,2}

MicroRNAs (miRNAs) in plants and animals function as post-transcriptional regulators of target genes, many of which are involved in multicellular development. miRNAs guide effector complexes to target mRNAs through base-pair complementarity, facilitating site-specific cleavage or translational repression. Biogenesis of miRNAs involves nucleolytic processing of a precursor transcript with extensive foldback structure. Here, we provide evidence that genes encoding miRNAs in plants originated by inverted duplication of target gene sequences. Several recently evolved genes encoding miRNAs in *Arabidopsis thaliana* and other small RNA-generating loci possess the hallmarks of inverted duplication events that formed the arms on each side of their respective foldback precursors. We propose a model for miRNA evolution that suggests a mechanism for *de novo* generation of new miRNA genes with unique target specificities.

miRNAs (~20–22 nucleotides) function as guides that direct effector complexes to specific mRNA targets, resulting in negative regulation through degradative pathways or cotranslational repression¹. miRNA biogenesis involves multistep processing of self-complementary (foldback) precursor transcripts. In animals, pri- (foldback excision) and pre-miRNA processing is catalyzed by the RNaseIII-like enzymes Drosha and Dicer^{2–4}; in plants, multiple steps may be catalyzed by Dicer-like1 (DCL1)^{5–7}. Although miRNAs have been found only in land plants and animals, most eukaryotes possess the RNA interference (RNAi) machinery through which miRNAs operate⁸. Formation of many genes encoding miRNAs in plants predates the split of monocots and dicots (~150 million years ago); many animal genes encoding miRNAs predate the divergence of metazoans (~600 million years ago)^{9,10}. There are no known orthologous miRNAs or common miRNA targets between plants and animals¹. Furthermore, several miRNA-controlled regulatory networks, such as those involving miR165/166, are found only in plant lineages^{11,12}. Plant miRNA targets typically contain single sites that are complementary to one miRNA family, whereas animal targets usually contain multiple, weakly complementary sites within 3' untranslated regions¹.

A. thaliana contains at least 22 miRNA families, most of which are conserved between monocots and dicots^{13,14}. Twelve of these families target mRNAs encoding transcription factors involved in development. Perturbation of miRNA biogenesis or targeting frequently leads to developmental defects, including alterations in leaf morphology, meristem identity, patterning and reproductive development^{15–21}. Several *A. thaliana* miRNAs affect hormone signaling pathways^{13,19}, and others regulate genes involved in stress responses^{13,22}. Two miRNAs

(miR162 and miR168) regulate genes required for miRNA biogenesis or activity^{23,24}. *A. thaliana* also contains several nonconserved miRNAs, such as miR161 and miR163. The nonconserved miRNAs are represented by single genes rather than multigene families¹³.

New gene regulatory networks can result from duplication of protein-coding sequences and regulatory elements^{25,26}. In this paper, we explore the possibility that genes encoding miRNAs in plants originated by duplication events from sequences in target gene families.

RESULTS

MIR161 and *MIR163* have extended similarity to target genes

Loci capable of forming a transcript that adopts an extended foldback structure can arise by inverted duplication events. If the originating sequence is a protein-coding gene, then the originating gene and closely related family members could be brought under negative regulation by RNAi through short interfering RNAs (siRNAs) spawned at the duplication locus⁸. If sequences at the duplication locus diverge under constraints to maintain a foldback structure and adapt to the miRNA biogenesis apparatus, then the new locus might evolve into a miRNA gene with specificity for one or more targets related to the founder gene. This is the inverted duplication hypothesis for evolution of genes encoding miRNAs.

The inverted duplication hypothesis predicts that the foldback arms of recently evolved miRNA genes might contain relatively long segments with similarity to target gene sequences. To investigate this possibility, we used the foldback sequences from 91 miRNA loci¹⁴ in FASTA searches against the *A. thaliana* gene database (Fig. 1). Complementarity or similarity to the ~21-nucleotide miRNA

¹Center for Gene Research and Biotechnology and ²Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331, USA. Correspondence should be addressed to J.C.C. (carrington@cgrb.oregonstate.edu).

Published online 21 November 2004; doi:10.1038/ng1478

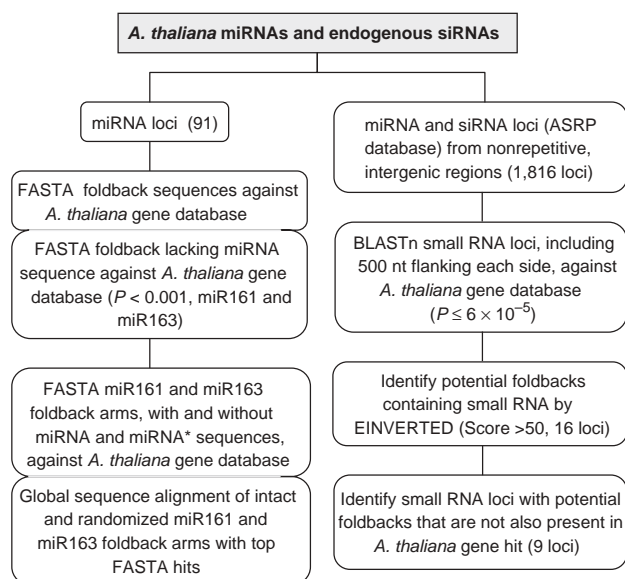


Figure 1 Flowchart for identification of miRNA foldbacks and endogenous small RNA loci with properties that are consistent with derivation by inverted duplication from protein coding genes. *A. thaliana* miRNA foldback sequences and endogenous siRNAs were as defined by the miRNA Registry¹⁴ and the ASRP database.

sequence alone was insufficient to detect a significant signal under the conditions used. Most miRNA foldbacks lacked extended similarity or complementarity to any gene sequences, including those from their respective target genes (Fig. 2a). But queries with foldback sequences from *MIR161* and *MIR163* resulted in significant hits to several genes, which corresponded to their respective target genes or closely related family members (Fig. 2a). We detected significant scores to the same genes even when the miRNA was computationally deleted from the foldback sequence (Fig. 2b,c). miR161 targets several mRNAs coding

for pentatricopeptide repeat proteins (PPRs), and miR163 was predicted to target several S-adenosylmethionine-dependent methyltransferases (SAMTs)^{5,27,28}. The *MIR161* locus is unusual in that it encodes overlapping miRNAs (miR161.1 and miR161.2) from a single precursor sequence (Fig. 3), comprising a contiguous, 29-nucleotide miR161.1-miR161.2 sequence containing complementarity to a contiguous sequence in each PPR target mRNA. miR163 is also unusual in that it is 24 nucleotides, rather than the more typical ~21 nucleotides, in length^{6,7,29} (Fig. 3).

We detected discrete segments of complementarity and similarity to target gene sequences in miRNA-containing and miRNA-complementing (miRNA*) arms, respectively, for *MIR161* and *MIR163* (Fig. 4a,b). The foldback arms of *MIR163* share significant similarity to sense and antisense polarities of a segment of three SAMT-like genes (Fig. 4b). The miR163 precursor also contains significant target sequence similarity in a region between the two foldback arms (Fig. 4b). We also detected segmental similarity between *MIR161* foldback arms and PPR target genes; the miR161* arm had the greatest similarity (Fig. 4a). To analyze these matches statistically, we used 1,000 randomized versions of each foldback arm from *MIR161* and *MIR163* in FASTA searches against the *A. thaliana* gene database. We also did this using arms lacking miRNA or miRNA-complementary sequences. We selected the top-scoring gene using each randomized sequence and subjected it to global sequence alignment with the corresponding query sequence. Using each intact *MIR161* and *MIR163* foldback arm, the top hit (always a target gene) scored 2.7–15.0 standard deviations (s.d.) ($P < 7 \times 10^{-3}$) higher than the mean of the top hits for the randomized sequences (Fig. 2c). Using foldback arms lacking the miR163 or miR163-complementary sequences, the top target gene hit scored at least 8.3 s.d. ($P < 1.3 \times 10^{-15}$) higher than the mean top score with the randomized arms. Among the arms lacking the miR161 or miR161-complementary sequences, top target gene scores were significantly higher ($P < 0.01$) than the randomized top score for only the miR161* arm lacking the miR161.1-complementary sequence (Fig. 2c). Although sequence similarity was clearly evident outside the miR161 and miR161-complementary regions of each *MIR161*

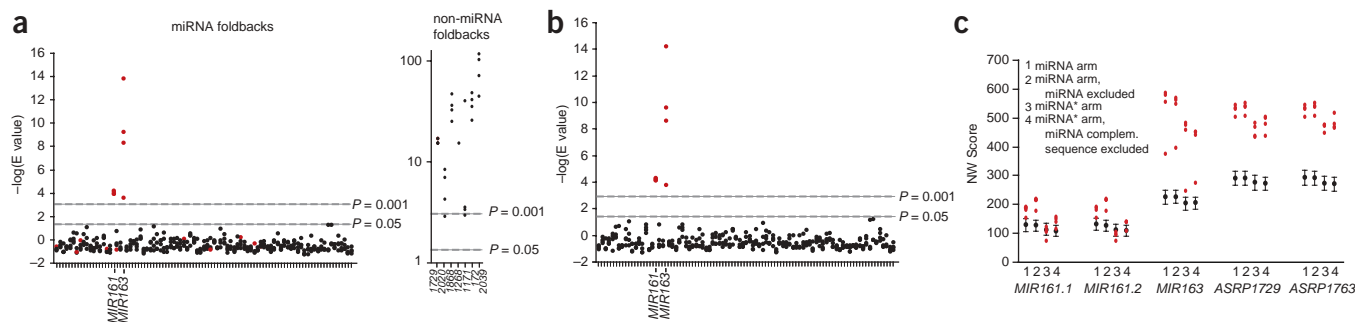


Figure 2 Computational analysis of miRNA and endogenous small RNA-generating foldback sequences. (a) On the left, *A. thaliana* gene matches in FASTA searches using foldback sequences from 91 miRNA genes (presented in numerical order from left to right in **Supplementary Table 1** online). The top four hits based on Expect value are presented. Genes corresponding to miRNA targets (or closely related target gene family members) or nontargets are indicated by red or black dots, respectively. Scores corresponding to $P = 0.05$ and $P = 0.001$ values are indicated. On the right, FASTA scores for the top four hits of predicted foldback sequences from seven small RNA-generating loci (**Table 1**). (b) Same analysis as in a, except that the miRNA sequence was computationally removed from each foldback sequence. (c) Needleman-Wunche (NW) scores of alignments between *MIR161*, *MIR163* and *ASRP1729* foldback arms and top four gene hits (red) detected by FASTA searches, and mean \pm s.d. of top gene hits detected using 1,000 shuffled sequences from each foldback arm. For *MIR163*, four foldback arm sequences were tested: (1) complete miRNA-containing foldback arm; (2) complete miRNA* arm; (3) miRNA arm with miR163 deleted; and (4) miRNA* arm with the miR163-complementary sequence deleted. For both *MIR161* and *ASRP1729*, two sets of analyses were done in which the adjacent or overlapping miRNA (miR161.1 and miR161.2) or small RNA (*ASRP1729* and *ASRP1763*) sequences were individually deleted.

foldback arm (Fig. 4a), the relative short length of the precursor segments after removal of miR161.1 or miR161.2 (or their complementary sequences) limited the statistical power of the methods used.

Inverted duplications at other small RNA-producing loci

The inverted duplication hypothesis predicts that a locus undergoes transitional evolutionary stages before acquiring miRNA-forming capacity. Initially, inverted duplication loci may contain segments with perfect or near-perfect repeats corresponding to sequences of protein-coding genes. If expressed, these transitional loci may yield heterogeneous siRNA populations just like transgenes engineered to express hairpin-forming transcripts^{30,31}. Biogenesis of these siRNAs may or may not require an RNA-dependent RNA polymerase, and they may form through the activity of DCL1 (as do most or all miRNAs) or another DCL protein associated with other classes of endogenous siRNAs³². We postulated that, if these transitional forms led to miRNA genes in the past, transitional forms with the potential to lead to miRNA genes in the future would be evident in the current *A. thaliana* genome. We devised a computational strategy to search for such loci (Fig. 1). We identified all known small RNA-generating loci from nonrepetitive intergenic sequences in the *Arabidopsis thaliana* Small RNA Project (ASRP) database and combined them with the miRNA loci (1,816 total loci). We used a genomic segment comprised of 500 nucleotides flanking both sides of each small RNA sequence, or up to the beginning of an adjacent gene, as a query sequence against the *Arabidopsis* gene database in BLASTn searches. Loci that hit one or more genes with a score that met or exceeded that of the *MIR161* locus (PPR target gene hit) were subjected to EINVERTED analysis to identify inverted duplications with the potential to form a foldback structure. We screened these loci manually to eliminate those in which the protein-coding gene(s) identified in the BLASTn search contained the same inverted duplication as the small RNA-generating locus.

We identified nine loci in the final set, two of which were *MIR161* and *MIR163* (Table 1). Each of the other seven contained inverted duplications with the potential to form foldback structures of ~195–2,684 nucleotides (Supplementary Fig. 1 online) and were represented by 1–20 small RNAs in the cloned sequence database (Table 1). The predicted foldback sequences and individual arms from each locus were used in FASTA searches against the *A. thaliana* gene database, revealing similarity or complementarity in each arm with at least one gene (Fig. 2a and Table 1). The ASRP2039 locus is part of a larger inverted duplication consisting of two annotated genes (At3g44570 and At3g44580). Part of the ASRP172 locus comprises a pseudogene (At4g04408) with similarity to histone H2A. The remaining five non-miRNA loci reside at nonannotated intergenic positions and contain similarity to genes encoding RAN1, flavin monooxygenase-like and F-box proteins, and proteins of unknown function (Table 1).

We subjected the predicted foldback structure containing ASRP1729 to more detailed analysis. The foldback region is represented by two tandem small RNAs in the database, ASRP1729 and ASRP1763 (Fig. 3). The entire length of each arm, approximately one-half of which is shown in Figure 4c, aligned with several genes coding for Divergent C1 (DC1) domain proteins. The DC1 family is represented by ~170 expressed or predicted *A. thaliana* genes of unknown function. Using the randomization technique described above, the statistical significance of similarity between each foldback arm and the most closely related DC1 domain gene sequences was very high ($P < 1.3 \times 10^{-15}$ with small RNAs either included or computationally deleted; Fig. 2c). Although relatively long (474 nucleotides) compared

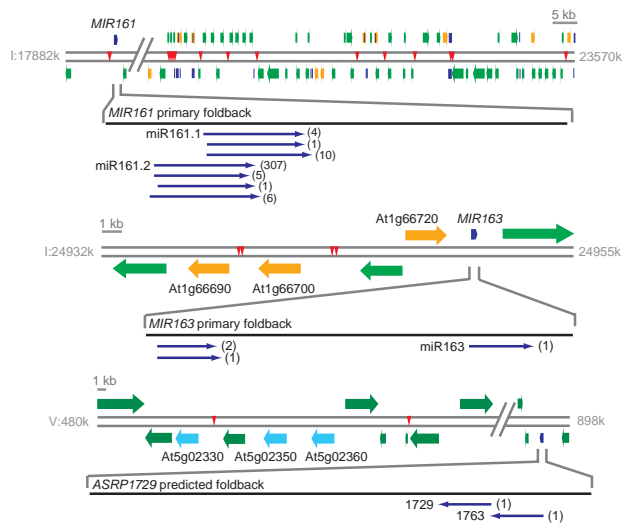


Figure 3 Genomic regions corresponding to *A. thaliana* *MIR161*, *MIR163* and the small RNA-generating locus *ASRP1729*. Target genes (orange arrows), nontarget genes (green arrows), retroelements (red triangles) and cloned miRNAs or small RNAs (blue lines) are shown. DC1 domain-containing genes (light blue arrows) related to *ASRP1729* are shown. The number of times each miRNA or small RNA was sequenced in the ASRP database is given in parentheses.

with *A. thaliana* miRNA foldback structures (139 ± 50 nucleotides), the ASRP1729 foldback contained mismatches and bulges similar to miRNA precursors (Fig. 4c).

Formation and function of miR161 and miR163

Because of the unusual properties and extended target gene similarity of *MIR161* and *MIR163*, we analyzed the biogenesis requirements and activity of miRNAs from each locus. Canonical miRNAs, such as miR169, generally require DCL1 (but not DCL2 or DCL3) and HEN1 but not RNA-dependent RNA polymerases RDR1, RDR2 and RDR6 (Fig. 5a)³². *trans*-acting siRNAs (such as ASRP255), a recently identified class of small RNAs that functions to suppress or degrade target mRNAs, require DCL1, HEN1 and RDR6 (Fig. 5a)^{33,34}. Chromatin-associated and perhaps other classes of endogenous siRNAs require DCL3 and RDR2 (ref. 32). *A. thaliana* mutants with defects in *DCL* and *RDR* genes are, therefore, useful to distinguish among different classes of small RNAs. Accumulation of miR163 was completely DCL1- and HEN1-dependent but was independent of DCL2, DCL3, RDR1, RDR2 and RDR6 (Fig. 5a). miR161.1 and miR161.2 had genetic requirements similar to those of miR163 and other miRNAs, but each was insensitive to the *dcl1-7* mutation (Fig. 5a). miR161.1 accumulation was lost in the *dcl1-9* mutant (Fig. 5b), however, indicative of a DCL1 requirement but allele-specific sensitivity to the two mutations³⁵. ASRP1729 was HEN1-dependent but was insensitive to individual mutations in each *dcl* mutant plant (Fig. 5a,b). This could reflect a biogenesis pathway that requires DCL4, which was not tested, or that involves redundant DCL activities. There was no evidence that accumulation of ASRP1729 small RNA required a specific RDR gene (Fig. 5a).

We obtained further evidence for miRNA function of miR161 and miR163 by analyzing several target mRNAs for each miRNA. Using a 5' RACE assay developed to map miRNA cleavage sites, which generally occur at a position 10 nucleotides from the 3' end of the miRNA-complementary sequence, we validated three PPR gene targets

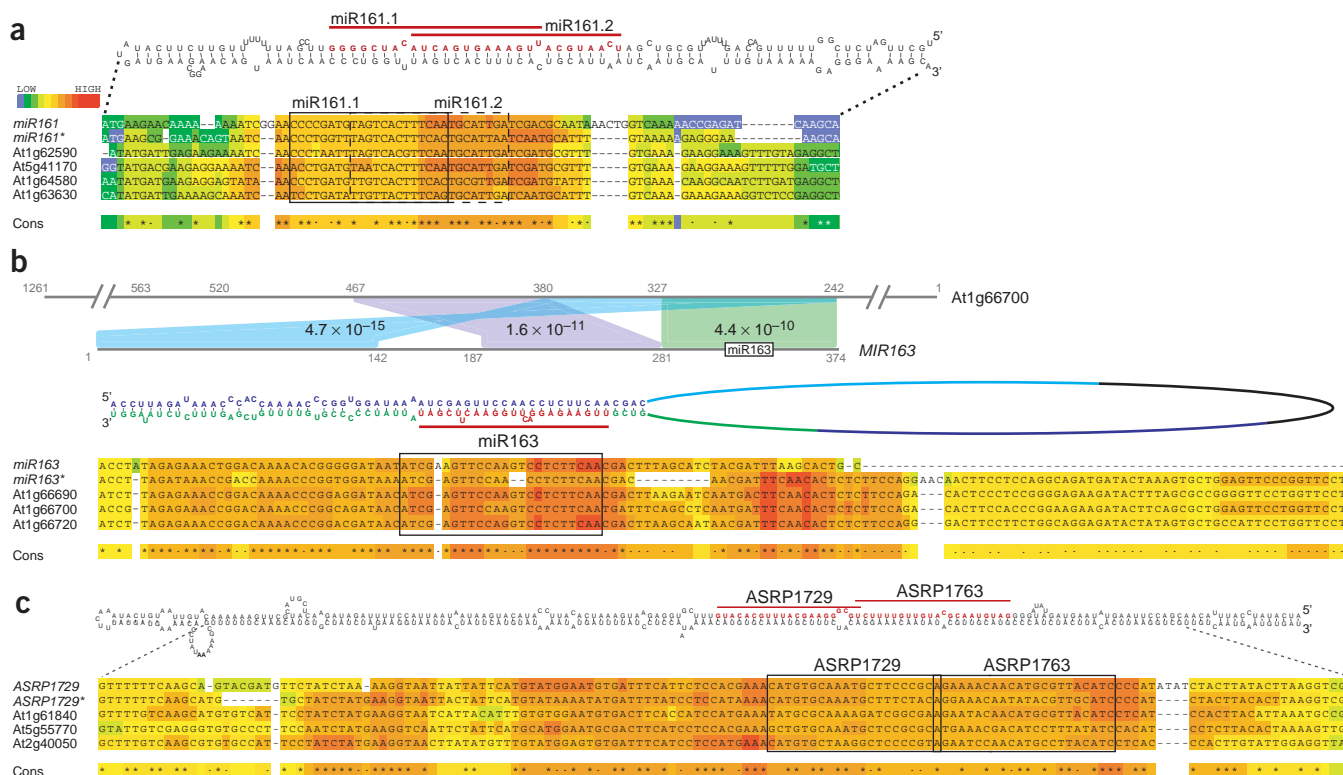


Figure 4 Similarity between foldback arms and protein-coding genes. Alignment of *MIR161* (a) and *MIR163* (b) foldback arm sequences and corresponding target genes, and *ASRP1729* foldback arms and most-closely related DC1 domain-containing genes (c). To visualize similarity, the sense orientation of each sequence was aligned with the actual miRNA* foldback arm and the reverse complement of the miRNA foldback arm. The color code for alignments represents the overall match quality at each position, as determined by T-Coffee. Color codes using T-Coffee indicate alignment quality in a regional context. Conserved positions across all sequences, and across target gene sequences and only one foldback arm, are indicated by asterisks and dots, respectively, in the consensus (Cons) bar. Positions corresponding to miRNA or miRNA-complementary sites are indicated. In b, a diagrammatic representation of *MIR163* sequences corresponding to duplication events is presented using the SAMT-like target gene At1g66700. The FASTA E value for each color-coded segment is shown.

(Fig. 5c). One of these (At1g06580) was validated previously²⁸. Based on the predominant position of cleavage, two PPR mRNAs were targeted by miR161.1 and one by miR161.2, indicating that both *MIR161*-derived species were functional. Each of four SAMT-like mRNAs were targeted by miR163 (Fig. 5c). Notably, we detected

cleavage at the canonical position relative to the 3' end of the complementary target site, despite the long length (24 nucleotides) of miR163. Several attempts to validate many DC1 domain-containing mRNAs (At1g61840, At1g53340, At1g44050, At2g13900, At2g13910, At2g02610, At2g02620, At2g02630, At5g02330, At5g02350, At3g26550,

Table 1 Small RNA-generating loci containing inverted gene duplications with similarity to protein-coding genes

| Small RNA locus | Small RNAs ^a | Foldback size (nt) | Gene | Domain family ^b | Similarity to <i>A. thaliana</i> genes | | | |
|-----------------|-------------------------|--------------------|-----------|----------------------------|--|---------------------|-----------------------|---------------------|
| | | | | | 5' arm | Strand ^c | 3' arm | Strand ^c |
| <i>MIR161</i> | 7 | 173 | At4g41170 | PPR | 8.1×10^{-4} | - | 1.3×10^{-5} | + |
| <i>MIR163</i> | 3 | 374 | At1g66700 | SAMT | 5.2×10^{-15} | + | 3.2×10^{-11} | - |
| <i>ASRP1729</i> | 2 | 474 | At1g61840 | DC1 | 3.2×10^{-12} | - | 2.0×10^{-12} | + |
| <i>ASRP2020</i> | 1 | 195 | At5g20010 | RAN | 9.2×10^{-8} | - | 7.9×10^{-10} | + |
| <i>ASRP1868</i> | 1 | 486 | At1g66290 | F-box | 3.9×10^{-36} | - | 1.2×10^{-45} | + |
| <i>ASRP1268</i> | 1 | 470 | At3g11000 | Unclassified | 1.0×10^{-8} | - | 8.0×10^{-4} | + |
| <i>ASRP1171</i> | 2 | 416 | At1g48910 | FMO | 2.0×10^{-32} | + | 1.0×10^{-73} | - |
| <i>ASRP172</i> | 1 | 810 | At3g54560 | H2A | 3.6×10^{-26} | - | 9.5×10^{-46} | + |
| <i>ASRP2039</i> | 20 | 2684 | At3g44570 | Unclassified | 0.0 ^d | + | 0.0 ^d | - |

^aUnique small RNA sequences in the ASRP database from each small RNA locus foldback. ^bFamily name abbreviations are as follows: PPR, pentatricopeptide repeat; SAMT, S-adenosyl-L-Met:salicylic acid carboxyl methyltransferase; DC1, Divergent C1; RAN, Ras-associated nuclear small GTP-binding protein; FMO, flavin monooxygenase; H2A, Histone 2A. ^cOrientation of the gene alignment relative to the small RNA locus. ^dFoldback contains FASTA match to annotated genes (At3g44570 and At3g44580).

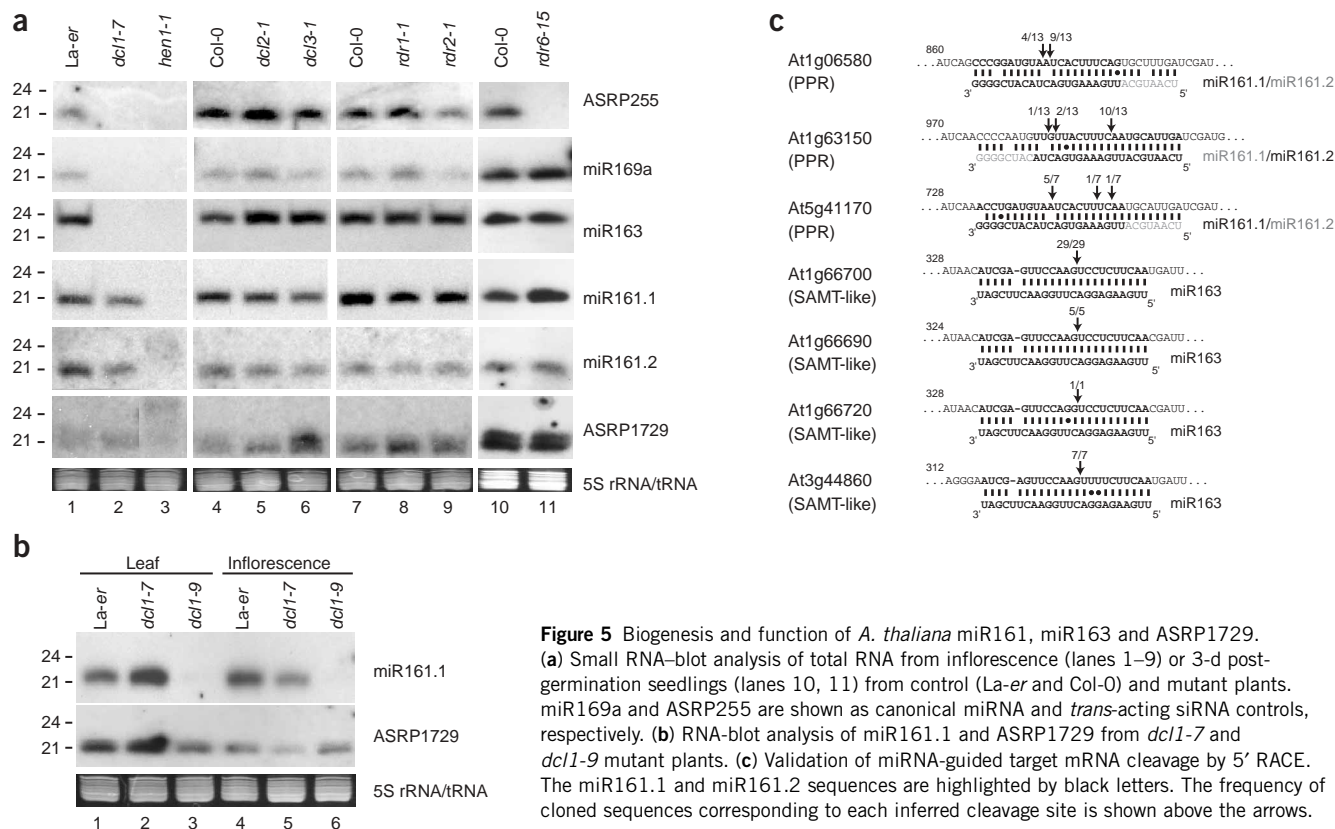


Figure 5 Biogenesis and function of *A. thaliana* miR161, miR163 and ASRP1729. (a) Small RNA-blot analysis of total RNA from inflorescence (lanes 1–9) or 3-d post-germination seedlings (lanes 10, 11) from control (*La-er* and *Col-0*) and mutant plants. miR169a and ASRP255 are shown as canonical miRNA and *trans*-acting siRNA controls, respectively. (b) RNA-blot analysis of miR161.1 and ASRP1729 from *dcl1-7* and *dcl1-9* mutant plants. (c) Validation of miRNA-guided target mRNA cleavage by 5' RACE. The miR161.1 and miR161.2 sequences are highlighted by black letters. The frequency of cloned sequences corresponding to each inferred cleavage site is shown above the arrows.

At1g62030, At1g47890, At1g53290 and At1g65680) as ASRP1729 or ASRP1763 targets were unsuccessful, although at least one of these mRNAs (At1g61840) was cleaved at a single position upstream from the putative target sites (data not shown). This could reflect the activity of another, as yet undetected, small RNA derived from the ASRP1729 foldback. Thus, despite some unusual properties, both miR161 and miR163 possess biogenesis and activity profiles that are typical of miRNAs. ASRP1729, however, has distinct biogenesis requirements and uncertain *trans*-active targeting functions.

Phylogenetics of *MIR161*, *MIR163* and *ASRP1729*

Most miRNA genes in *A. thaliana* are not genetically linked to their respective target genes. In contrast, *MIR163* is located immediately adjacent to a cluster of three SAMT-like target genes (Fig. 3b). *MIR161* is not tightly linked to target genes but is ~5.7 Mbp away from a region containing a high density of predicted and validated PPR target genes (Fig. 3). The *ASRP1729* locus resides ~0.4 Mbp away from a cluster of closely related DC1 domain-containing genes with high similarity (Fig. 3). The PPR, SAMT-like and DC1 domain-containing gene clusters resulted from relatively recent duplication events and were inferred to be members of rapidly evolving families. The proximity of miRNA genes and other foldback-encoding loci to expanding target families is suggestive of an evolutionary relationship.

To test the relationships between *MIR161* and *MIR163* and their respective target gene families, we carried out phylogenetic reconstructions (Bayesian and maximum parsimony methods³⁶) using each foldback arm as an independent taxon in a gene set that included validated targets, predicted targets and the 23 (*MIR161*) or 18 (*MIR163*) most closely related target family members. The topology of both trees supported monophyly of each miRNA foldback arm and

target genes (Fig. 6a,b). *MIR161* arms each clustered in a well-supported clade containing the 3 validated and 16 predicted miR161 targets (Fig. 6a). The *MIR163* foldback arms clustered tightly with the three validated SAMT-like target genes located adjacent to the *MIR163* locus and formed a larger clade consisting exclusively of all six validated or predicted miR163 targets (Fig. 6b).

We also generated phylogenetic trees using *ASRP1729* foldback arms and 39 DC1 domain-containing genes. Both *ASRP1729* foldback arms clustered with 25 annotated members of the DC1 domain family in a strongly supported clade (Fig. 6c). Given the consensus properties of known small RNA–target interactions that trigger mRNA cleavage^{13,27,37,38}, most of the genes in this clade are potential targets for RNAi triggered by the *ASRP1729*-derived small RNAs.

DISCUSSION

We conclude that *A. thaliana* *MIR161* and *MIR163* genes evolved relatively recently by inverted duplication events associated with active expansion of target gene families and then adapted to the miRNA biogenesis apparatus. The proposed evolutionary pathway includes transitional gene forms that probably spawned siRNAs and that resemble several inverted duplication loci derived from protein-coding genes (such as *ASRP1729*) evident in the *A. thaliana* genome today. We propose a general mechanism to spawn and evolve miRNA genes with unique target specificities (Fig. 7). Inverted duplication events from a founder gene (step 1) result in head-to-head or tail-to-tail orientations of complete or partial gene sequences. Duplication may include the founder gene promoter or may result in capture of a new promoter. The inverted duplication can occur directly from a genomic sequence or, potentially, during integration of a pseudogene-like sequence after reverse transcription. The new locus can even form

through juxtaposition of two closely related sequences from different members of a gene family. Transcription of the nascent locus yields foldback transcripts that potentially serve as DCL substrates for siRNA biogenesis, bringing the founder gene and closely related family members under control of RNAi at the post-transcriptional or chromatin levels⁸. Given that miRNA target sites generally occur outside of family-defining domains^{15,18,19,27}, we suggest that duplication events yielding foldbacks from highly conserved domains are under strong negative selection. Limited sequence divergence at the inverted dupli-

cation locus (step 2), under constraints to maintain both the foldback character and recognition by a DCL activity, generates one or more specific siRNA populations, such as those from the *ASRP1729* locus. Adaptation of the foldback transcript to the miRNA biogenesis pathway involving DCL1 (step 3) results in formation of a uniform miRNA population. Additional sequence divergence (step 4), under foldback and DCL1-recognition constraints, occurs to the point that only the miRNA or miRNA-complementary sequences resemble the originating gene sequence. Duplication of the miRNA locus (step 5) results in a

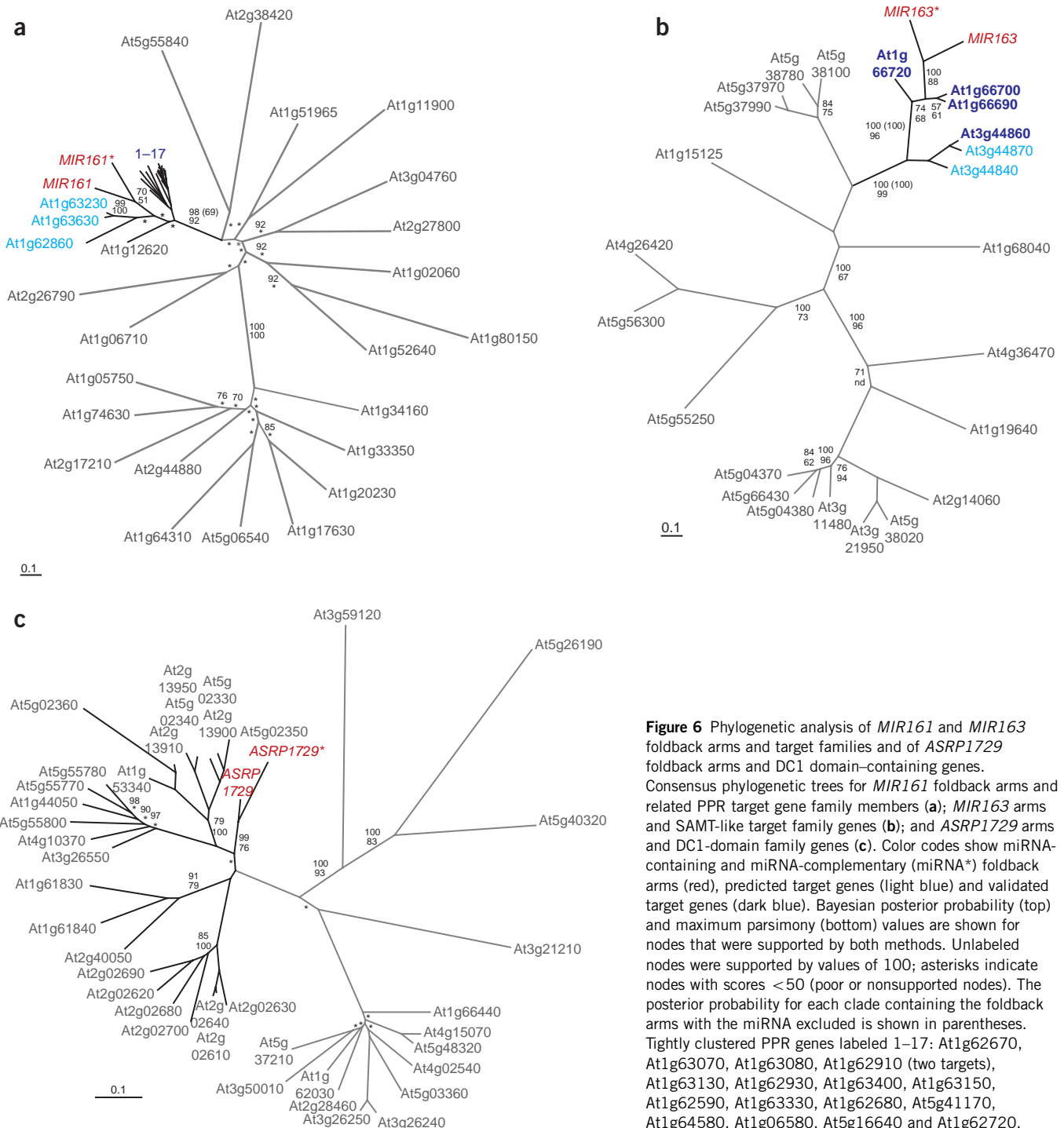
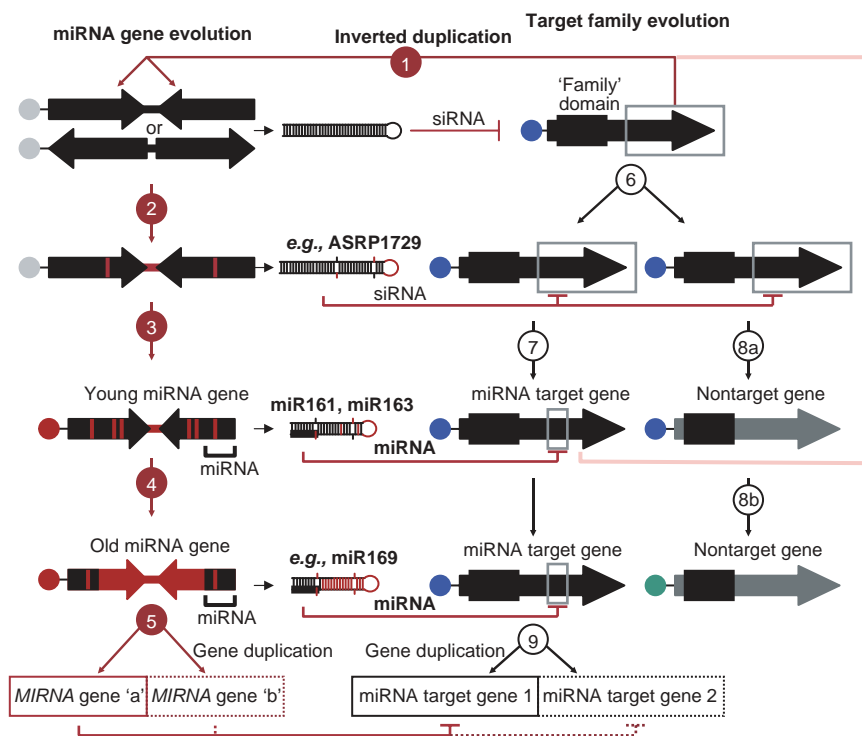


Figure 6 Phylogenetic analysis of *MIR161* and *MIR163* foldback arms and target families and of *ASRP1729* foldback arms and DC1 domain-containing genes. Consensus phylogenetic trees for *MIR161* foldback arms and related PPR target gene family members (**a**); *MIR163* arms and SAMT-like target family genes and DC1-domain family genes (**b**); and *ASRP1729* arms and DC1-domain family genes (**c**). Color codes show miRNA-containing and miRNA-complementary (miRNA*) foldback arms (red), predicted target genes (light blue) and validated target genes (dark blue). Bayesian posterior probability (top) and maximum parsimony (bottom) values are shown for nodes that were supported by both methods. Unlabeled nodes were supported by values of 100; asterisks indicate nodes with scores < 50 (poor or nonsupported nodes). The posterior probability for each clade containing the foldback arms with the miRNA excluded is shown in parentheses. Tightly clustered PPR genes labeled 1–17: At1g62670, At1g63070, At1g63080, At1g62910 (two targets), At1g63130, At1g62930, At1g63400, At1g63150, At1g62590, At1g63330, At1g62680, At5g41170, At1g64580, At1g06580, At5g16640 and At1g62720.

Figure 7 Inverted duplication model for miRNA gene evolution in plants. Red arrows and step labels indicate miRNA gene evolution events. Black arrows and clear step labels indicate target family evolution events. Sequences in proposed transitional genes and miRNA genes with sequence similarity or complementarity to target gene sequences are indicated by black shading. Sequences that are unique or divergent relative to target gene sequences are indicated in red. Putative or known foldback structures from each transitional locus or miRNA gene are shown.



multigene miRNA family, perhaps leading to miRNA target specificity after further sequence divergence. Most modern miRNA genes of *A. thaliana* are members of multigene families³⁹ and lack extended target gene similarity outside of the miRNA or miRNA-complementary sequences. We propose that *MIR161* and *MIR163* are young genes that have progressed only to step three.

The model is incomplete without consideration of target-gene family evolution (Fig. 7). Most miRNA target genes in plants are subsets of larger gene families^{9,13,15,18,19,27,40}. Target-family gene duplication (step 6) provides the pool from which regulatory diversification emerges. After formation of an siRNA- or young miRNA-generating locus (steps 2 or 3), retention (step 7) or loss (step 8a) of small RNA-complementary target sites among family members leads to differential post-transcriptional regulation. This may be accompanied by changes in transcriptional control elements (step 8b), leading to further regulatory diversification. Further duplication and subsequent diversification of miRNA target genes (step 9) leads to specialized regulation by distinct miRNA family members. Thus, new regulatory networks might emerge through a series of duplication events followed by retention or loss of sequence complementarity between regulator (miRNA) and target genes. The evolutionary force driving these events is probably selective advantage gained by differential regulation of target family members²⁶. Application of RNA-based control may be particularly well-suited for genes conferring cell fate properties or responses to stress^{13,22}.

Seven *A. thaliana* small RNA-generating loci, including *ASRP1729*, possess the properties predicted for transitional forms resulting after step 2 in the model (Fig. 7). We do not consider these to be miRNA genes because the small RNA biogenesis requirements are not clear. Also, the functionality of these loci as negative regulators of related protein-coding genes has not been established. Further, the ultimate evolution of these loci toward miRNA gene status cannot be determined at this point. We propose, however, that these and similar loci comprise an evolutionary reservoir from which to sample RNA-based elements that confer advantageous regulatory properties to target family members⁴¹. This model does not explain how loci that generate *trans*-acting siRNA evolved^{33,34}. Evidence is lacking for transcripts with miRNA-like foldback structure arising from loci that generate *trans*-acting siRNA. It is possible that these loci yield transcripts that are adapted as RDR6 templates rather than direct DCL substrates^{33,34}.

The miRNA evolution model offers a mechanism for *de novo* generation of new, RNA-based regulatory genes from protein-coding genes. If all plant miRNA genes arose through this mechanism, it may

explain why there are no common miRNAs or miRNA-regulated targets between plants and animals. Lack of orthologous miRNA genes between kingdoms is expected if plant miRNA genes arose *de novo* after the split of animal-fungi and plant lineages. But we cannot rule out the possibility that other evolutionary pathways, though currently unknown or unexplored, contributed to the miRNA gene class in plants. This model does not necessarily apply to animal miRNA genes, which encode foldback structures that are too short to search for evidence of derivation from target genes. Bartel and Chen⁴² proposed that animal miRNAs function primarily as redundant modulators to fine-tune expression of thousands of genes through combinatorial interactions within 3' untranslated regions. Thus, new animal target genes are more likely to come under the control of miRNAs using existing miRNAs through gain-of-interaction events, such as target gene recombination or duplication of the 3' untranslated region sequence, rather than by *de novo* miRNA formation events.

Finally, given that most plant miRNAs regulate genes belonging to families with roles in development, miRNA-directed regulation may have provided selective advantages for evolution of a multicellular body plan. Recent evidence suggests that organism complexity arises primarily from application of new regulatory control over duplicated genes rather than by invention of new activities⁴³. The coincident expansion of gene families controlling development and miRNA-based control mechanisms may have had a profound influence on the independent evolution of multicellularity in plants and animals¹¹.

METHODS

Identification of small RNA loci from inverted duplications of protein-coding genes. We identified matches between small RNA loci in the ASRP database and protein-coding genes (The Institute for Genome Research AGI gene database, version 3.0) by BLAST searches using segments containing 500 nucleotides flanking both sides of the small RNA. We included only small RNAs from nonrepetitive, intergenic regions in the search. We treated each small RNA sequence as an independent locus, regardless of overlap among related

sequences. We omitted segments of annotated genes within 500 nucleotides of small RNA loci from the analysis. We analyzed small RNA loci with gene matches with an expect value $E \leq 6 \times 10^{-5}$ for potential to form a foldback structure using EINVERTED⁴⁴. We manually inspected loci with an EINVERTED score > 50 for the presence of an inverted repeat that included the small RNA sequence. We eliminated small RNA-generating loci that had BLAST hits to protein-coding genes containing the same or similar inverted duplication. We predicted foldback structures using RNAfold⁴⁵.

Computational analysis of miRNA foldback and target gene sequences. We used FASTA to search The Institute for Genome Research AGI transcript database (version 4.0) for genes with similarity to a reference set of 91 miRNA precursor foldback sequences (Supplementary Table 1 online). Searches used a 5–4 matrix. We also did a second search using the foldback with the mature miRNA sequence excluded. We analyzed further those foldback sequences (with miRNA sequence computationally deleted) that hit gene sequences with an expect value $E < 0.001$ by FASTA searches against the transcript database using each foldback arm independently. We calculated global alignment scores using the foldback arms and gene sequences with NEEDLE using the Needleman-Wunche alignment algorithm⁴⁴. Gene sequences from At5g41170, At1g64590, At1g62590 and At1g63630 were aligned to the *MIR161* foldback arms; sequences from At1g66700, At1g66690, At1g66720 and At3g44840 were aligned to *MIR163* arms; and sequences from At1g61830, At2g13900, At5g02340 and At2g13910 were aligned to *ASRP1729* arms. We repeated this analysis with foldback arms lacking miRNA or miRNA-complementary sequences. To evaluate the significance of the Needleman-Wunche score, we created 1,000 shuffled sequences for each miRNA foldback segment using SHUFFLESEQ⁴⁴. We used FASTA to find the top-scoring gene hit (based on expect value) for each shuffled sequence. We calculated Needleman-Wunche scores for each shuffled sequence and corresponding top gene hit and used them to determine a mean and s.d. for shuffled foldback and target sequence alignments. We calculated the probability of a random match⁴⁶.

miRNA and small RNA-blot analysis. We isolated total RNA from independent pools of inflorescence or 3-d-post-germination seedlings using Trizol reagent (Invitrogen) and purified it with an RNA/DNA Midi Kit (Qiagen). We resolved total RNA (7.5 μ g) on a 17% polyacrylamide-urea gel, blotted it to a nylon membrane and probed it with a ³²P end-labeled oligonucleotide probe³⁰. *A. thaliana* mutants containing *dcl1-7*, *dcl1-9*, *dcl2-1*, *dcl3-1*, *rdrl-1*, *rdrl-2* and *hen1-1* were described previously^{5,32,35}. The *rdrl-15* allele (SAIL_617) contains a T-DNA insertion at a position 312 nucleotides beyond the start codon of At3g49500.

5' RACE analysis of miRNA target genes. We mapped cleavage sites in the mRNAs of miRNA target genes using a modified 5' RACE procedure as described previously⁴⁰. We designed gene-specific primers ~ 500 nucleotides downstream of the predicted miRNA target site. We cloned purified PCR products into pGEM-T Easy (Promega) and sequenced them.

Phylogenetic reconstructions. We aligned the amino acid sequences of proteins from genes used in phylogeny reconstruction using T-Coffee⁴⁷ and then converted them to the corresponding nucleotide sequence using TRANALIGN⁴⁴. We then aligned miRNA precursor arms to the prealigned protein coding sequences using T-Coffee and excluded the ambiguous regions. We carried out Bayesian phylogenetic analyses with MrBayes v3.0B4 (ref. 48). Each codon of each miRNA or target family was assigned its own model of evolution using likelihood ratio test implemented in Modeltest⁴⁹. In each case, we used the GTR + γ model for individual positions. Analyses were done using four chains and run for 1,000,000 generations. Trees were sampled every tenth generation and the resulting 100,000 trees were plotted against their likelihoods to discard all trees before the convergence point in the burn-in phase. A majority-rule consensus was generated with the remaining 90,000 trees for each gene to calculate the posterior probability. The maximum parsimony analyses were also done for each gene using PAUP⁵⁰ with the following options: 100 replicates of random sequence addition, TBR (Tree bisection-reconnection) branch swapping and Multrees in effect. Relative support for the resulting tree was determined by 1,000 bootstrap replications and the corresponding bootstrap values were presented in the Bayesian consensus trees with the posterior

probability. The topology of both Bayesian and maximum parsimony trees had similar topology for all gene families.

URLs. The ASRP database is available at <http://asrp.cgrb.oregonstate.edu/>.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank S. Givan, D. Smith and C. Sullivan for assistance and advice with computational resources; L. Johansen for initial propagation of the *rdrl-15* mutant; and S. Poethig and H. Vaucheret for discussions about *trans*-acting siRNAs and the suggestion to analyze miR161 in multiple *dcl1* mutated alleles. This work was supported by grants from the US National Science Foundation, the US National Institutes of Health and the United States Department of Agriculture.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 17 August; accepted 26 October 2004

Published online at <http://www.nature.com/naturegenetics/>

- Bartel, D. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
- Ketting, R.F. *et al.* Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev.* **15**, 2654–2659 (2001).
- Hutvagner, G. *et al.* A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834–838 (2001).
- Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).
- Park, W., Li, J., Song, R., Messing, J. & Chen, X. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12**, 1484–1495 (2002).
- Kurihara, Y. & Watanabe, Y. *Arabidopsis* micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc. Natl. Acad. Sci. USA* **101**, 12753–12758 (2004).
- Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B. & Bartel, D.P. MicroRNAs in plants. *Genes Dev.* **16**, 1616–1626 (2002).
- Finnegan, E.J. & Matzke, M.A. The small RNA world. *J. Cell Sci.* **116**, 4689–4693 (2003).
- Floyd, S.K. & Bowman, J.L. Gene regulation: ancient microRNA target sequences in plants. *Nature* **428**, 485–486 (2004).
- Pasquinelli, A.E. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86–89 (2000).
- Meyerowitz, E.M. Plants compared to animals: the broadest comparative study of development. *Science* **295**, 1482–1485 (2002).
- Poethig, R.S. Life with 25,000 genes. *Genome Res.* **11**, 313–316 (2001).
- Jones-Rhoades, M.W. & Bartel, D.P. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell* **14**, 787–799 (2004).
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S.R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).
- Palatnik, J.F. *et al.* Control of leaf morphogenesis by microRNAs. *Nature* **425**, 257–263 (2003).
- Aukerman, M.J. & Sakai, H. Regulation of flowering time and floral organ identity by a microRNA and its APETALA2-like target genes. *Plant Cell* **15**, 2730–2741 (2003).
- Chen, X. A microRNA as a translational repressor of APETALA2 in *Arabidopsis* flower development. *Science* **303**, 2022–2025 (2004).
- Mallory, A.C., Dugas, D.V., Bartel, D.P. & Bartel, B. MicroRNA regulation of NAC-domain targets is required for proper formation and separation of adjacent embryonic, vegetative, and floral organs. *Curr. Biol.* **14**, 1035–106 (2004).
- Achard, P., Herr, A., Baulcombe, D.C. & Harberd, N.P. Modulation of floral development by a gibberellin-regulated microRNA. *Development* **131**, 3357–3365 (2004).
- Emery, J.F. *et al.* Radial patterning of *Arabidopsis* shoots by class III HD-ZIP and KANADI genes. *Curr. Biol.* **13**, 1768–1774 (2003).
- Laufs, P., Peaucelle, A., Morin, H. & Traas, J. MicroRNA regulation of the CUC genes is required for boundary size control in *Arabidopsis* meristems. *Development* **131**, 4311–4322 (2004).
- Sunkar, R. & Zhu, J.K. Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. *Plant Cell* **16**, 2001–2019 (2004).
- Xie, Z., Kasschau, K.D. & Carrington, J.C. Negative feedback regulation of Dicer-Like1 in *Arabidopsis* by microRNA-guided mRNA degradation. *Curr. Biol.* **13**, 784–789 (2003).
- Vaucheret, H., Vazquez, F., Crete, P. & Bartel, D.P. 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 (2004).
- Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).

26. Teichmann, S.A. & Babu, M.M. Gene regulatory network growth by duplication. *Nat. Genet.* **36**, 492–496 (2004).
27. Rhoades, M.W. *et al.* Prediction of plant microRNA targets. *Cell* **110**, 513–520 (2002).
28. Vazquez, F., Gasciolli, V., Crete, P. & Vaucheret, H. The nuclear dsRNA binding protein HYL1 is required for microRNA accumulation and plant development, but not posttranscriptional transgene silencing. *Curr. Biol.* **14**, 346–351 (2004).
29. Dunoyer, P., Lecellier, C.H., Parizotto, E.A., Himber, C. & Voinnet, O. Probing the microRNA and small interfering RNA pathways with virus-encoded suppressors of RNA silencing. *Plant Cell* **16**, 1235–1250 (2004).
30. Llave, C., Kasschau, K.D., Rector, M.A. & Carrington, J.C. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**, 1605–1619 (2002).
31. Papp, I. *et al.* Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant Physiol.* **132**, 1382–1390 (2003).
32. Xie, Z. *et al.* Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol.* **2**, E104 (2004).
33. Peragine, A., Yoshikawa, M., Wu, G., Albrecht, H.L. & Poethig, R.S. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in *Arabidopsis*. *Genes Dev.* **18**, 2368–2379 (2004).
34. Vazquez, F. *et al.* Endogenous trans-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs. *Mol. Cell* **16**, 69–79 (2004).
35. Schauer, S.E., Jacobsen, S.E., Meinke, D.W. & Ray, A. *DICER-LIKE1*: blind men and elephants in *Arabidopsis* development. *Trends Plant Sci.* **7**, 487–491 (2002).
36. Holder, M. & Lewis, P.O. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**, 275–284 (2003).
37. Mallory, A.C. *et al.* MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. *EMBO J.* **23**, 3356–3364 (2004).
38. Kasschau, K.D. *et al.* P1/HC-Pro, a viral suppressor of RNA silencing, interferes with *Arabidopsis* development and miRNA function. *Dev. Cell* **4**, 205–217 (2003).
39. Bartel, B. & Bartel, D.P. MicroRNAs: at the root of plant development? *Plant Physiol.* **132**, 709–717 (2003).
40. Llave, C., Xie, Z., Kasschau, K.D. & Carrington, J.C. Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* **297**, 2053–2056 (2002).
41. Herbert, A. The four Rs of RNA-directed evolution. *Nat. Genet.* **36**, 19–25 (2004).
42. Bartel, D.P. & Chen, C.Z. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat. Rev. Genet.* **5**, 396–400 (2004).
43. Hurles, M. Gene duplication: the genomic trade in spare parts. *PLoS Biol.* **2**, E206 (2004).
44. Rice, P., Longden, I. & Bleasby, A. EMBOS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
45. Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431 (2003).
46. Lin, J.-T. Alternatives to Hamaker's approximations to the cumulative normal distribution and its inverse. *Statistician* **37**, 413–414 (1988).
47. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
48. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
49. Posada, D. & Crandall, K.A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818 (1998).
50. Swofford, D. *PAUP*: Phylogenetic analysis using parsimony (*and other methods)* 4.0b10 edn. (Sinauer, Sunderland, Massachusetts, 2002).