Comparative Analysis of the MIR319a MicroRNA Locus in Arabidopsis and Related Brassicaceae

Norman Warthmann,*1,2 Sandip Das,*1,2 Christa Lanz,† and Detlef Weigel*†

*Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany; and †Genome Center, Max Planck Institute for Developmental Biology, Tübingen, Germany

MicroRNAs (miRNAs) are important regulators of gene expression in multicellular organisms. Yet, little is known about their molecular evolution. The 20- to 22-nt long miRNAs are processed in plants from foldbacks that are a few hundred base pairs in size. Often, these foldbacks are embedded in much larger precursor transcripts. To investigate functional constraints on sequence evolution of miRNA precursor genes, we have studied sequence variation in the precursor of miR319a, MIR319a, between species from the Brassicaceae. We compared the genomic context in Arabidopsis thaliana, Arabidopsis halleri, and Capsella rubella, using bacterial artificial chromosome clones, and analyzed precursor sequences obtained by polymerase chain reaction from 13 additional species. Phylogenetic shadowing identifies a conserved motif around the transcription start site, which we demonstrate to be functionally important. We further assessed the functionality of MIR319a orthologs from several Brassicaceae species in A. thaliana. The ortholog from kale (Brassica oleracea var. acephala) was found to be largely inactive, at least partially due to mutations in the miRNA itself, but experimental evidence suggests that loss of miR319a function is compensated by other members of the miR319 family. More broadly, we find that the foldback diverges less rapidly than the remainder of the primary transcript. To understand the molecular evolution of miRNA genes, investigations at different levels of phylogenetic divergence are required.

Introduction

MicroRNAs (miRNAs) are small RNAs around 21 nt in length and have widespread roles as posttranscriptional regulators of plant and animal physiology and development. They can affect target genes through a variety of mechanisms, including transcript cleavage and translational repression. The mature miRNAs are derived from larger precursor transcripts that contain a self-complementary foldback structure known as pri-miRNA. The precursors are processed by ribonuclease III type enzymes of the Drosha and Dicer family, first releasing the pri-miRNA and then a short duplex with a 2-nt overhang at the 3' end, consisting of the miRNA and its complement, miRNA*. The primary precursor transcripts are capped and polyadenylated and can be spliced, in agreement with them being transcribed by RNA polymerase II (Kim 2005; Jones-Rhoades et al. 2006).

Many plant miRNAs are encoded by gene families, and the mature miRNAs in turn often have multiple targets with very similar complementary motifs in their miRNAs. The relationship between multiple miRNAs and multiple targets might explain that there is apparently very little co-evolution between targets and miRNAs, causing many miRNAs to be astonishingly highly conserved in sequence even over large evolutionary distances (Floyd and Bowman 2004; Arazì et al. 2005; Axtell and Bartel 2005; Talmor-Neiman et al. 2006; Axtell et al. 2007; Fattash et al. 2007). In contrast, there is relatively little sequence conservation apart from the miRNA and miRNA*, especially outside the pri-miRNA. (Aukerman and Sakai 2003; Palatnik et al. 2003; Gustafson et al. 2005; Nikovics et al. 2006).

Because miRNA precursors do not encode proteins, they are subject to different evolutionary constraints, in terms of both spatial and temporal patterns of sequence evolution. In the case of conserved miRNAs, the sequence of the miRNA itself evolves very slowly, whereas the surrounding foldback is often not conserved at all between, for example, monocotyledonous and dicotyledonous plants (Jones-Rhoades et al. 2006). Recently, important progress has been made in understanding the constraints on evolution of the foldback by comparing miRNA genes from 12 Drosophila species (Stark et al. 2007), but an important distinction between animals and plants is the very similar size of animal miRNA foldbacks, about 80 bp, whereas plant foldbacks are much more variable. In plants, there is limited understanding of adaptive evolutionary constraints among closely related taxa. As is the case in animals, little attention has been given to the entire precursor (Guddeti et al. 2005; Maher et al. 2006). Comparing sequences of miRNA precursors within a species and between closely related species should thus help to determine patterns of molecular evolution and the timescales at which different aspects of the evolution of miRNA precursors are best investigated.

We have chosen to study miR319a, the first plant miRNA identified by forward genetics (Palatnik et al. 2003). The miR319a target genes that encode TCP transcription factors, which have important roles in controlling leaf growth. Reduced miR319 activity and thus increased TCP function limit vegetative growth and are in the most extreme cases lethal, whereas in plants with increased miR319 expression and reduced TCP activity, there is excessive cell division and leaf overgrowth (Nath et al. 2003; Palatnik et al. 2003). Some of these phenotypes are reminiscent of what one sees in certain forms of domesticated Brassica oleracea varieties, and it is conceivable that modulation of miR319 activity has contributed to the variation in leaf morphology in B. oleracea. The miR319 has been detected in diverse flowering plants and also in the nonvascular land plant Physcomitrella patens, a moss...
of genome sequences completed to date, the genome of black cottonwood, *Populus trichocarpa*, is the closest to that of *Arabidopsis thaliana* (Tuskan et al. 2006), but we found little obvious sequence conservation outside of the *mir319a* foldbacks in this species. We therefore assessed intraspecific variation in *A. thaliana*, isolated, and compared orthologs of the *mir319a* precursor gene, *MIR319a*, from closely related species. We began with completely assembled bacterial artificial chromosome (BAC) sequences from 2 close relatives, *Arabidopsis halleri* and *Capsella rubella*, and subsequently isolated sequences of additional orthologs from 13 Brassicaceae. The activity of 3 orthologs from *Arabidopsis lyrata*, *Sibara virginica*, and *B. oleracea*, and of a paralog, *MIR319c* from *B. oleracea*, was tested by misexpression in *A. thaliana*.

**Materials and Methods**

**Oligonucleotide Sequences**

See supplementary table 1 (Supplementary Material online) for sequences of oligonucleotides used for polymerase chain reaction (PCR) amplification, mutagenesis, and RNA blot hybridization.

**BAC Library Screening and Sequencing**

To identify BACs that contain *MIR319a* orthologous sequences, high-density BAC filters of genomic DNA libraries of *A. halleri* and *C. rubella* were probed with an α-P32-dCTP-labeled PCR fragment of the *A. thaliana* *MIR319a* locus amplified using the primer pair N-0547 and N-0548. Filter hybridizations were carried out overnight at 42°C. Filters were washed once with 2× standard saline citrate (SSC), 0.1% sodium dodecyl sulfate (SDS) at 65°C for 20 min followed by 2 washes with 0.2× SSC, 0.1% SDS at 65°C for 20 min and then exposed to Kodak BioMax MS films with 2 washes with 0.1% SDS at 65°C for 20 min. Primer pairs were initially selected based on sequence alignments between *A. thaliana*, *A. lyrata*, *Arabidopsis cebennensis*, *A. halleri*, and *C. rubella*; G-3389 and G-3390. For *B. oleracea* *MIR319a*, identity of clones was confirmed by misexpression in *A. thaliana*. For intraspecific comparison of the entire locus, a 3.7-kb fragment surrounding the *MIR319a* from 19 different *A. thaliana* accessions (supplementary table 2, Supplementary Material online) was amplified with primers G-1703 and G-2536 using the *Pfu* Turbo Polymerase (Stratagene, Amsterdam, the Netherlands), a high-fidelity enzyme, blunted, cloned into pBluescript, and completely sequenced on both strands with a total of 22 primers.

For the amplification of *MIR319a* orthologs from other species, primer pairs were initially selected based on *A. thaliana* (Col-0) genomic sequence: N-0545 or N-0547 combined with N-0550. Alternatively, consensus primers were designed based on sequence alignments between *A. thaliana*, *A. lyrata*, *Arabidopsis cebennensis*, *A. halleri*, and *C. rubella*; G-3389 and G-3390. For *B. oleracea* *MIR319a*, information retrieved from the Brassica genome gateway (http://brassica.bbsrc.ac.uk; GenBank accession numbers BH572910, BH007837, BH666711, BH724635, BH714840, BH007836, BH720638, and BZ515979) was used to design specific oligonucleotide primers: G-1836 and G-1837. Primers with sequences and their purpose are listed in supplementary table 1 (Supplementary Material online). For sequencing, the PCR products were gel purified using QiaQuick Gel columns (Qiagen) and cloned into either pGEM-T Easy (Promega, Mannheim, Germany) or pBluescript. Two or 3 clones from each species were sequenced. Identity of the species under study was confirmed by amplifying and sequencing the internal transcribed spacers of rDNA (ITS) with the primer pair N-0471/N-0472.

**Sequence Analysis**

To annotate the assembled BAC sequences, global alignments of BAC sequence contigs from *A. halleri* and *C. rubella* to *A. thaliana* genomic sequences were performed with AVID (Bray et al. 2003). Alignments were visualized using VISTA (VISualization Tool for Alignments, Frazer et al. 2004). Repetitive sequences were identified with RepeatMasker, open-3.0 (Smit et al. 1996–2004) and by comparison to the plant repeat database of The Institute of Genomic Research (TIGR v2). Tandem repeats were found by tandem repeats finder (Benson 1999) and by using EMBOSS (Rice et al. 2000). Gene prediction was based on the output from AVID and VISTA, GenScan (Burge and Karlin 1997), GeneMark.hmm (Lomsadze et al. 2005), and FGENESH (http://www.softberry.com) were used to predict putative novel open reading frames and validated using TBlastN against the non-redundant nucleotide collection of the National Center for Biotechnology Information (NCBI).

*MIR319a* orthologs (table 1, fig. 2B–D) were aligned using MUSCLE (v3.41) (Edgar 2004). Maximum likelihood phylogenies (heuristic search tree) from this multiple sequence alignment (1,666 bp) of the complete sequences isolated were generated with PAUP (Swofford 1993) using the substitution model HKY85. Bootstrap confidence values were obtained by 100 replications. The phylogeny was generated to illustrate orthology of the isolated sequences to *A. thaliana MIR319a*, and nodes that are supported by high bootstrap values agree with the species phylogeny.
Sequence divergence was analyzed with DnaSP version 4.10.9 (Rozas et al. 2003). Phylogenetic shadowing to identify conserved sequences by comparing closely related species was performed using eShadow (Ovcharenko et al. 2004).

For intraspecies comparisons of the MIR319a locus, sequences from 19 A. thaliana accessions (Supplementary table 2, Supplementary Material online) were aligned using MUSCLE (v3.41) (Edgar 2004), and all analyses (π, Tajima’s D, Fay and Wu’s H, and the Hudson-Kreitman-Aguade [HKA] test) were performed using DnaSP version 4.10.9 (Rozas et al. 2003). The Fay and Wu tests were performed with the most conservative assumption of no recombination, and the significance of Fay and Wu’s H was assessed by coalescent simulations with 10,000 replicates. The species A. lyrata belongs to the closest known relatives of A. thaliana (Bailey et al. 2006) and was chosen as outgroup. The corresponding genomic sequence from Arabidopsis lyrata ssp. lyrata was retrieved and assembled from the trace archive in NCBI and manually aligned to the multiple sequence alignment of the 19 A. thaliana accessions.

The HKA test was used to test whether the rate of evolution between the upstream region (most likely containing the promoter) and the miRNA precursor is different using all polymorphic positions in the intergenic region between At4g23710 and the transcription start of MIR319a and all polymorphic positions found in splice variant 1.

For the pairwise interspecies comparisons (fig. 1D), the best local alignment of the entire 3.7 kb surrounding the jaw locus from A. thaliana (Col-0) to the corresponding sequences from A. lyrata ssp. lyrata (see above), A. halleri (BAC), C. rubella (BAC), and Brassica rapa ssp. pekinensis (GenBank accession number AC189494, clone KBrH054J11; GenBank accession number AJ856769) was assessed by coalescent simulations with 10,000 replicates. The species A. lyrata belongs to the closest known relatives of A. thaliana (Bailey et al. 2006) and was chosen as outgroup. The corresponding genomic sequence from Arabidopsis lyrata ssp. lyrata was retrieved and assembled from the trace archive in NCBI and manually aligned to the multiple sequence alignment of the 19 A. thaliana accessions.

The HKA test was used to test whether the rate of evolution between the upstream region (most likely containing the promoter) and the miRNA precursor is different using all polymorphic positions in the intergenic region between At4g23710 and the transcription start of MIR319a and all polymorphic positions found in splice variant 1.

For the pairwise interspecies comparisons (fig. 1D), the best local alignment of the entire 3.7 kb surrounding the jaw locus from A. thaliana (Col-0) to the corresponding sequences from A. lyrata ssp. lyrata (see above), A. halleri (BAC), C. rubella (BAC), and Brassica rapa ssp. pekinensis (GenBank accession number AC189494, clone KBrH054J11; GenBank accession number AJ856769) was assessed by coalescent simulations with 10,000 replicates. The species A. lyrata belongs to the closest known relatives of A. thaliana (Bailey et al. 2006) and was chosen as outgroup. The corresponding genomic sequence from Arabidopsis lyrata ssp. lyrata was retrieved and assembled from the trace archive in NCBI and manually aligned to the multiple sequence alignment of the 19 A. thaliana accessions.

The HKA test was used to test whether the rate of evolution between the upstream region (most likely containing the promoter) and the miRNA precursor is different using all polymorphic positions in the intergenic region between At4g23710 and the transcription start of MIR319a and all polymorphic positions found in splice variant 1.

For the pairwise interspecies comparisons (fig. 1D), the best local alignment of the entire 3.7 kb surrounding the jaw locus from A. thaliana (Col-0) to the corresponding sequences from A. lyrata ssp. lyrata (see above), A. halleri (BAC), C. rubella (BAC), and Brassica rapa ssp. pekinensis (GenBank accession number AC189494, clone KBrH054J11; GenBank accession number AJ856769) was assessed by coalescent simulations with 10,000 replicates. The species A. lyrata belongs to the closest known relatives of A. thaliana (Bailey et al. 2006) and was chosen as outgroup. The corresponding genomic sequence from Arabidopsis lyrata ssp. lyrata was retrieved and assembled from the trace archive in NCBI and manually aligned to the multiple sequence alignment of the 19 A. thaliana accessions.

The HKA test was used to test whether the rate of evolution between the upstream region (most likely containing the promoter) and the miRNA precursor is different using all polymorphic positions in the intergenic region between At4g23710 and the transcription start of MIR319a and all polymorphic positions found in splice variant 1.

For the pairwise interspecies comparisons (fig. 1D), the best local alignment of the entire 3.7 kb surrounding the jaw locus from A. thaliana (Col-0) to the corresponding sequences from A. lyrata ssp. lyrata (see above), A. halleri (BAC), C. rubella (BAC), and Brassica rapa ssp. pekinensis (GenBank accession number AC189494, clone KBrH054J11; GenBank accession number AJ856769) was assessed by coalescent simulations with 10,000 replicates. The species A. lyrata belongs to the closest known relatives of A. thaliana (Bailey et al. 2006) and was chosen as outgroup. The corresponding genomic sequence from Arabidopsis lyrata ssp. lyrata was retrieved and assembled from the trace archive in NCBI and manually aligned to the multiple sequence alignment of the 19 A. thaliana accessions.

The HKA test was used to test whether the rate of evolution between the upstream region (most likely containing the promoter) and the miRNA precursor is different using all polymorphic positions in the intergenic region between At4g23710 and the transcription start of MIR319a and all polymorphic positions found in splice variant 1.

For the pairwise interspecies comparisons (fig. 1D), the best local alignment of the entire 3.7 kb surrounding the jaw locus from A. thaliana (Col-0) to the corresponding sequences from A. lyrata ssp. lyrata (see above), A. halleri (BAC), C. rubella (BAC), and Brassica rapa ssp. pekinensis (GenBank accession number AC189494, clone KBrH054J11; GenBank accession number AJ856769) was assessed by coalescent simulations with 10,000 replicates. The species A. lyrata belongs to the closest known relatives of A. thaliana (Bailey et al. 2006) and was chosen as outgroup. The corresponding genomic sequence from Arabidopsis lyrata ssp. lyrata was retrieved and assembled from the trace archive in NCBI and manually aligned to the multiple sequence alignment of the 19 A. thaliana accessions.

The HKA test was used to test whether the rate of evolution between the upstream region (most likely containing the promoter) and the miRNA precursor is different using all polymorphic positions in the intergenic region between At4g23710 and the transcription start of MIR319a and all polymorphic positions found in splice variant 1.

For the pairwise interspecies comparisons (fig. 1D), the best local alignment of the entire 3.7 kb surrounding the jaw locus from A. thaliana (Col-0) to the corresponding sequences from A. lyrata ssp. lyrata (see above), A. halleri (BAC), C. rubella (BAC), and Brassica rapa ssp. pekinensis (GenBank accession number AC189494, clone KBrH054J11; GenBank accession number AJ856769) was assessed by coalescent simulations with 10,000 replicates. The species A. lyrata belongs to the closest known relatives of A. thaliana (Bailey et al. 2006) and was chosen as outgroup. The corresponding genomic sequence from Arabidopsis lyrata ssp. lyrata was retrieved and assembled from the trace archive in NCBI and manually aligned to the multiple sequence alignment of the 19 A. thaliana accessions.
We had mapped the transcription start of MIR319a in A. thaliana and found the corresponding sequence surrounding the transcription start to be highly conserved in most Brassicaceae species analyzed. To test its functional relevance for expression of MIR319a, constructs pNW8 and pNW9 were produced for reporter studies, where construct pNW9 lacks half of the conserved sequence. The intergenic region between the At4g23710 and the MIR319a precursor was amplified from A. thaliana genomic DNA with Pfu Turbo Polymerase (Stratagene) using the sense primer G-1825 together with the antisense primers G-1823 and G-1824, which gave rise to a 1,206-bp and 1,328-bp fragment, respectively, and subsequent cloning into pRita (Gleave 1992) in front of the β-glucuronidase (GUS) reporter gene using introduced XhoI and EcoRI restriction sites. A NotI fragment was then transferred into the binary vector pMLBart (Gleave 1992).

All clones for transformation were sequence verified, and the A. thaliana accession Col-0 was transformed using floral dip into an Agrobacterium tumefaciens suspension (Clough and Bent 1998). Plants were grown in either long days (23°C, 16 h light) or short days (23°C, 8 h light).

Small RNA Isolation and Blot Analysis

Successful expression of miR319a and miR319c from transgenes was assessed by RNA blot hybridization using a 21-bp probe complementary in sequence to miR319a. The RNA blot (fig. 4) was made from inflorescences of plants grown in long-day conditions. Two inflorescence clusters containing stage 1–12 flowers from 3 plants each were pooled, and total RNA was isolated as described earlier (Smyth et al. 1990; Palatnik et al. 2003). Five micrograms of RNA were resolved on a 17% polyacrylamide gel electrophoresis under denaturing conditions and transferred to a charged nylon membrane (Nytran SuPer Charge, Whatmann/Schleicher&Schuell, Dassel, Germany) by semidry blotting (Llave et al. 2002). The blot was hybridized with 5 pmol of radioactively end-labeled locked nucleic acid (LNA) oligo probe (Exiqon, Vedbæck, Denmark) (Vester and Wengel 2004) complementary to the mature A. thaliana miR319a sequence (supplementary table 1, Supplementary Material online) in PerfectHyb Plus (Sigma, Taukirchen, Germany) hybridization buffer at 42°C overnight. After hybridization, the blot was briefly rinsed with 2× SSC, 0.2% SDS, washed for 25 min with 1× SSC, 0.1% SDS at 50°C, and exposed to Kodak BioMax MS film for 18 h with 2 intensifying screens at ~80°C.

Rapid Amplification of cDNA Ends

To characterize the complete pri-miRNA transcript of MIR319a, rapid amplification of cDNA ends (RACE) of
both ends from *A. thaliana* MIR319a primary transcripts was performed with the SMART RACE cDNA Amplification Kit (Clontech/Takara Bio Europe, Saint-Germain-en-Laye, France) using total RNA from young flowers. The gene-specific primers for the 3′ RACE were G-0570 (primary PCR), N-0549, G-1138, and G-1243 (secondary, nested PCR) and for the 5′ RACE G-0571 (primary PCR), G-1681, G-1682, and G-1683 (secondary, nested PCR). RACE–PCR products were gel purified, cloned into pGEM-T Easy, and 40 clones were sequenced for each end.

**Results**

**MIR319 Family in Arabidopsis thaliana**

To begin to understand the potential constraints on the evolution of transcribed and untranscribed sequences, we first characterized the MIR319a primary transcript in detail; we mapped the 5′ and 3′ ends by RACE, using RNA from flowers. Transcription initiation was found to occur about 500-bp upstream of the foldback (fig. 1A), at position 12,352,498 of chromosome IV (TAIR7 genome release). Approximately 30-bp upstream of the transcription initiation site, a motif resembling a TATA Box (TTATAAA) is present, in agreement with this miRNA gene, like others in plants, being transcribed by RNA polymerase II (Aukerman and Sakai 2003; Palatnik et al. 2003; Lee et al. 2004). The 3′ RACE products indicated alternative splicing downstream of the foldback, giving rise to 2 different pri-precursors. A 1,403 base transcript (GenBank accession number AY922325) is intron less, whereas the primary transcript giving rise to a mature 1,295 base form (GenBank accession number AY922324) contains an 1,180 base intron (fig. 1A), the ends of which conform to the splice site consensus GU...AG also found in protein coding genes.

Apart from MIR319a, the *A. thaliana* genome contains 2 genes giving rise to identical or very similar mature miRNAs, MIR319b and MIR319c (Palatnik et al. 2003; Sunkar and Zhu 2004). The MIR319a (At4g23713) is flanked by 2 protein-coding genes, At4g23710 (encoding vacuolar adenosine triphosphate synthase subunit G2) and At4g23720 (encoding an expressed protein) (Weigel et al. 2000) (fig. 1A). The homolog of MIR319a on chromosome V, MIR319b (At5g41663), between protein-coding genes At5g41660 (encoding an expressed protein) and At5g41670 (encoding 6-phosphogluconate dehydrogenase) (Palatnik et al. 2003) can give rise to the same mature miRNA as MIR319a. Inspection of a database of segmental duplications (Blanc et al. 2003), which predate the split of the lineages being analyzed in this study, indicated that MIR319a and MIR319b are located in an old duplication, from At4g23470 to At4g24140 on chromosome IV and At5g41390 to At5g41900 on chromosome V. Of the approximately 60 genes in these 2 blocks, only 6 including the 2 MIR319 loci have been retained (fig. 2A).

MIR319c has only recently been annotated in the latest *A. thaliana* genome release (TAIR7 genome release) as At2g40805. It is found between protein-coding genes At2g40800 (encoding an expressed protein) and At2g40810 (encoding ATG18a/WD-40 repeat family protein). It gives rise to a mature miRNA, miR319c, that differs from miR319a/b by 1 nucleotide (Palatnik et al. 2003, 2007; Sunkar and Zhu 2004; Gustafson et al. 2005). The MIR319c region is contained in a large segmental duplication (Blanc et al. 2003) that occurred at the base of the Brassicaceae, covering 300 genes and extending over about 4 Mb on chromosomes II and III. The region between the flanking genes At2g40800 and At2g40810 spans about 5.5 kb and includes the MIR319c locus. The corresponding region between At3g56430 and At3g56440 on chromosome III is shorter than 600 bp and does not contain a MIR319c paralog, indicating that this copy of MIR319c has been lost.

**MIR319a Sequence Polymorphisms in Arabidopsis thaliana**

We began our comparative study of MIR319a sequences by analyzing a 3.7-kb fragment of the locus from 19 accessions of *A. thaliana* (supplementary table 2, Supplementary Material online). The analyzed region extended from position 12,351,365 (end of the 3′ untranslated region [UTR] of At4g23710) to 12,355,026 of chromosome IV (TAIR7 genome release). The locus could be amplified from all accessions studied, and excluding the primer matching sites, a total of 3,669 sites (3,571 without gaps) were analyzed. Sixty-three SNPs and indels were found. Of these, 38 occur in more than 1 ecotype. Within the 190 bp of the foldback, there are only 2 nucleotide changes (both singletons), without any changes in the miRNA itself. In addition, 300 bp around the transcription start are devoid of any polymorphisms.

The average number of pairwise differences per site (π) was 0.004. Compared with a genome-wide survey (Nordborg et al. 2005), the observed value of π is higher than what is reported for exons but lower than the genome-wide average for intergenic, intron, and UTR sequences. Whereas Tajima’s D (Tajima 1989) was −0.92 over the entire 3.7-kb region, it was not significantly different from 0 (P > 0.1). Tajima’s D is expected to be 0 under the neutral model and negative values reflect an excess of low-frequency mutations, which can result either from selective sweeps or from demographic processes. The genome-wide distribution of Tajima’s D in *A. thaliana* is known to be shifted toward negative values, indicating that demographic processes such as population expansion are the cause for the overall excess of low-frequency mutations in this species (Nordborg et al. 2005; Schmid et al. 2005). Tajima’s D at MIR319a falls within the range of genome-wide expectations in *A. thaliana* (Nordborg et al. 2005). Therefore, MIR319a is unlikely to be a recent target of directional selection in *A. thaliana*. Also not significant was a Fay and Wu test (Fay and Wu 2000) with the sister species *A. lyrata* as outgroup (H = −7.98, P = 0.098). However, it is worth noting that a sliding window analysis of Fay and Wu’s H along the entire locus revealed that most of the high-frequency derived mutations are found in the intergenic region upstream of MIR319a (fig. 1C), which could reflect recent cis-regulatory changes. We performed an HKA test (Hudson et al. 1987) using all polymorphic positions in the upstream versus the pre-miRNA region (splice variant 1) but detected no significant difference in the evolutionary dynamics of the upstream intergenic region and the pre-miRNA (χ² = 0.679, P = 0.41).
The *P. trichocarpa* genome is compared with *A. thaliana*, the most closely related genome sequence completed today. *Populus trichocarpa* has undergone an extra genome duplication (Tuskan et al. 2006) and by consulting an alignment of both genomes (http://pipeline.lbl.gov/), 2 foldbacks were easily identified on linkage groups I and III of *P. trichocarpa*, with 77.4% and 74.7% nt identity to the *A. thaliana* stem-loop sequence of MIR319a. There was little obvious sequence conservation outside of the foldback, indicating that this taxon was too distant to provide information on the evolution of surrounding sequences. We therefore chose to isolate and shotgun sequence BACs containing the MIR319a locus from *A. halleri*, a species in the same genus as *A. thaliana*, and *C. rubella*, a close relative in the Brassicaceae (table 1). The assembled BAC sequences were 131 kb and 147 kb long, respectively. The finished and annotated sequences have been deposited in GenBank (accession numbers EF197846 and EF197847).

The 131-kb contig from *A. halleri* (GC content: 35.6%) corresponds to a syntenic 76.3-kb stretch of chromosome IV of *A. thaliana*, covering 18 genes, At4g23690 to At4g23860 (supplementary fig. 2A and table 3, Supplementary Material online). The 147-kb contig from *C. rubella* (GC content: 35.4%) corresponds to a syntenic 134-kb region in *A. thaliana*, covering 34 genes, At4g23420 to At4g23720 (supplementary fig. 2A and table 3, Supplementary Material online). A total of 30 of these genes were detected in the *C. rubella* contig, 3 genes, At4g23430, At4g23550, and At4g23580, are missing entirely and 1, At4g23510, is partially deleted. The assembly indicates that there is a large inversion about 70 kb upstream of the MIR319a locus. The average sequence identity in coding and noncoding regions is 94.7% and 80.7% between *A. thaliana* and *A. halleri*, respectively. It is lower between *A. thaliana* and *C. rubella*, with 91% and 74.0%, respectively. Both *A. halleri* and *C. rubella* feature several repeats and transposons (supplementary table 4, Supplementary Material online). In addition, there were several short local duplications, either in tandem or in inverted orientation.

The region surrounding the MIR319a locus, extending from At4g23710 to At4g23720, was analyzed in more detail using VISTA to detect highly conserved regions (fig. 1B). In between the coding regions of At4g23710 and At4g23720 orthologs, an intergenic region with 85% nt differences was detected.
identity extends over 1.8 kb between A. thaliana and A. halleri and harbors MIR319a. Between A. thaliana and C. rubella, a 400-bp intergenic region surrounding MIR319a has a sequence identity of 79.5%. The foldback of MIR319a, which is approximately 180 bp long, is even more highly conserved across all 3 species with 99% and 96% identity between A. thaliana–A. halleri and A. thaliana–C. rubella, respectively (fig. 1B, circled).

MIR319a in Other Brassicaceae

The genomic sequences for the MIR319a locus from the sister species A. lyrata and A. cebennensis were easily obtained by PCR amplification using primers based on the A. thaliana sequence, and the sequence for B. rapa ssp. pekinensis was retrieved from GenBank (accession number AC189494; position 53,548–53,723 bp). Degenerate oligonucleotide primers were designed based on conserved sequences between A. thaliana, A. lyrata, A. cebennensis, A. halleri, and C. rubella and used to isolate MIR319a sequences from 10 additional Brassicaceae (table 1). Cloning of MIR319a from B. oleracea, the most distantly related species examined, was facilitated by the availability of genomic shotgun sequences (http://brassica.bbsrc.ac.uk). The length of the sequences recovered by PCR, ranged from 1,130 bp in Rorippa indica and Nasturtium officinale to 1,871 bp in B. oleracea var. acephala (kale), with the MIR319a foldback being located approximately in the middle. Taken together, apart from A. thaliana we obtained MIR319a sequences from 16 additional Brassicaceae species (table 1).

Phylogenetic analysis suggested that all the sequences isolated from other Brassicaceae are orthologous to A. thaliana MIR319a, because they are more closely related to MIR319a than to MIR319b, which arose from a duplication preceding the diversification of Brassicaceae but apparently after divergence from the lineage leading to Populus, which experienced an independent duplication (fig. 2A and B). Within the Brassicaceae, both, DnaSP and eShadow (Rozas et al. 2003; Ovcharenko et al. 2004) identified 2 regions of reduced sequence divergence. The least divergent region is around 190 bp long and comprises the pre-miRNA foldback itself (fig. 2C). An individual comparison of species for which the entire upstream intergenic and transcribed regions were available with A. thaliana confirmed the foldback as the least variable segment (fig. 1D). However, the degree of conservation decays with increasing phylogenetic distance, and in the pairwise comparison between A. thaliana and B. rapa, strong conservation is only detected for MIR319a sequences (position 53,548–53,723 bp), showed that the mature miR319a sequence from B. rapa is identical to that of A. thaliana, indicating that the B. oleracea SNP in miR319a is not shared with other members of the genus. Approximately 470 bp upstream of the foldback, there is a second region of reduced sequence divergence among the Brassicaceae, surrounding the transcription initiation site (figs. 1B and 2C and D). However, it is apparently not conserved in B. oleracea.

Functional Analysis of MIR319 Orthologs

To determine whether the MIR319a genes of other species could be properly processed and give rise to
functional miR319a, we constitutively expressed the foldbacks from 3 species, A. lyrata, S. virginica, and B. oleracea var. acephala, in A. thaliana plants under control of the viral 35S promoter, which confers strong and near-ubiquitous expression (Odell et al. 1985) (see supplementary table 5 [Supplementary Material online] for details on the overexpression constructs). The sequences outside the foldbacks were variable in length but always comprised sequences known to be part of the primary transcript in A. thaliana. The length of sequence outside the foldback has little, if any effect on the phenotype of miRNA overexpression in A. thaliana transgenic plants (Parizotto et al. 2004; Schwab et al. 2005).

Overexpression of A. thaliana miR319a in A. thaliana causes cotyledon epinasty, crinkled leaves, and siliques, due to miR319-guided degradation of mRNAs encoding a series of related TCP transcription factor genes (fig. 3B) (Palatnik et al. 2003, 2007). The same phenotypes were seen when miR319a from A. lyrata and S. virginica were overexpressed (fig. 3C and D). In contrast, transgenic A. thaliana plants transformed with the B. oleracea var. acephala overexpression construct, which contains a divergent miR319a sequence (fig. 3A, supplementary fig. 1 [Supplementary Material online]), had normal cotyledons, rosette leaves, and siliques (fig. 3E). We made the same observation with plants overexpressing an almost identical construct but with the miR319a sequence mutated so that it conforms to the canonical sequence (fig. 3F). A substantial fraction of T1 lines of both constructs (60% for the unmutated and 33% for the mutated version) had reduced seed set due to impaired pollen production (fig. 3E and F).

Overexpression experiments and target predictions indicate that the 3 miR319 isoforms in A. thaliana, miR319a, miR319b, and miR319c, have similar roles (Palatnik et al. 2007). Examining various Brassica expressed sequence tag (EST) databases, we found that 3 ESTs had been isolated with high similarity to MIR319c in A. thaliana but none for MIR319a or MIR319b. To determine whether B. oleracea miR319c could potentially substitute for miR319a/b, we overexpressed the foldback of B. oleracea MIR319c in A. thaliana (supplementary fig. 1, Supplementary Material online). These transgenic plants had crinkly leaves and epinastic cotyledons (fig. 3H), with the phenotypic effects being more pronounced than those of A. thaliana MIR319c (Palatnik et al. 2007) (fig. 3G).

To ensure that the phenotypic defects were indeed due to miR319 overexpression, we isolated small RNAs and analyzed them by blot hybridization. High levels of processed miR319a were detected in plants, carrying the overexpression constructs from A. thaliana, S. virginica, and A. lyrata (fig. 4). No overexpressed mature miRNA was seen in plants transformed with the B. oleracea miR319a constructs, suggesting that the pri-miRNA is not efficiently processed or less likely expressed. Mutating the miR319a sequence in this construct so that it conforms to the canonical sequence did not improve its expression (fig. 4). In contrast, miR319c from B. oleracea is efficiently processed in A. thaliana, consistent with its ability to efficiently cause the typical miR319a gain-of-function defects (figs. 3H and 4).

Detection of a Regulatory Element by Phylogenetic Shadowing

The transcription initiation site, as mapped in A. thaliana, coincides with the second most conserved region 470 bp upstream of the foldback (figs. 1A and D and 2C and D). To determine the possible importance of this region for MIR319a expression, we generated 2 promoter fusions to the GUS reporter (pNW8 and pNW9). Both constructs comprise all intergenic sequence up to the next annotated transcription start. Reporter pNW8 includes 143 bp of the transcribed region, whereas pNW9 includes only 21 bp (fig. 5).

Transgenic seedlings and adult plants were harvested at different stages and assayed for reporter activity. GUS activity could be easily detected in several tissues of plants carrying the longer reporter construct pNW8, including the hypocotyl of germinating seedlings and flowers (fig. 5). In contrast, plants with the shorter reporter construct pNW9, which lacks half of the conserved motif around the transcription start site, did not exhibit any GUS activity (fig. 5).

Discussion

Plant miRNAs are often processed from larger precursor transcripts; yet, there is little sequence conservation outside the miRNA across larger evolutionary distances. This is also true for MIR319a, which is found in both A. thaliana and P. trichocarpa. An analysis of the orthologous genomic regions showed that sequence similarity is restricted to the foldback region of MIR319a. To begin to better understand the evolutionary dynamics of MIR319a, we therefore chose to focus on shorter evolutionary distances that may not yet have led to an elimination of most sequence similarity outside the foldback. To this end, we studied sequence divergence of MIR319a within A. thaliana as well as between A. thaliana and close relatives from the same family.

These analyses confirmed the same trend as seen in sequence comparisons across larger evolutionary distances,
namely that sequences diverge more rapidly outside the foldback. A broadly similar finding has recently been reported for the MIR156b and MIR156c loci in cereals and especially rice (Wang et al. 2007). Phylogenetic shadowing, the multiple alignment of orthologous sequence from many closely related species (Boffelli et al. 2003; Hong et al. 2003), did, however, allow us to identify a region under stronger evolutionary constraints, overlapping the transcription start site of MIR319a. Reporter experiments demonstrated that this sequence contains important regulatory information for MIR319a transcription because removing the conserved sequences downstream of the transcription initiation site eliminated reporter gene activity. Whether these sequences are important for transcription initiation, or act as transcriptional enhancers, is not known; however, it is noteworthy that this sequence has diverged in B. oleracea.

Outside the MIR319a foldback, sequence divergence of MIR319a transcribed sequences within A. thaliana was intermediate between what has been reported for coding and noncoding regions (Nordborg et al. 2005). Across other Brassicaceae, we found that there is more sequence divergence in the middle of the MIR319a foldback, where the sequences form an unpaired loop. Again, this observation is consistent with the pattern of sequence divergence seen with more distantly related species, including other dicots and even monocots (Palatnik et al. 2003, 2007).

We analyzed the biological activity of MIR319a homologs by overexpression in A. thaliana. Overexpression of MIR319a from A. lyrata and S. virginica, which are predicted to produce a mature miR319a identical to that of A. thaliana, caused very similar phenotypes to overexpression of MIR319a from A. thaliana. In contrast, attempted overexpression of MIR319a from B. oleracea, where the mature miRNA differs in a single nucleotide from miR319a of other Brassicaceae, failed. This did not appear to be simply due to the change in the mature miRNA sequence itself because correcting it to conform to the canonical miR319a sequence did not improve accumulation of the miRNA, suggesting that additional changes in MIR319a contribute to the failure in processing.

Many miRNAs are encoded by small gene families (Jones-Rhoades et al. 2006); this is also the case for miR319. Overexpression experiments with a MIR319c ortholog from B. oleracea indicated that miR319c, despite small sequence differences, has very similar activity as miR319a in this species. Interestingly, in A. thaliana, overexpression of MIR319c does not seem to be as effective as overexpression of MIR319a, most likely because miR319c does not accumulate as efficiently as miR319a or miR319b (Palatnik et al. 2007). We therefore propose that the MIR319a gene of B. oleracea is becoming a pseudogene and that its function is assumed by either MIR319b (for which we have no expression evidence based on public databases) or MIR319c.

In summary, our work has illuminated aspects of the short-term evolutionary trajectory of a plant miRNA gene. The impending completion of the entire genome sequence of 2 close A. thaliana relatives, A. lyrata and C. rubella, promises to provide additional important insights into plant miRNA evolution.

Supplementary Material

Supplementary figures 1 and 2 and tables 1–5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/). GenBank accession numbers are as follows: AY775957–AY775959, AY775961–AY775968, AY775970, AY775971, AY775974, AY775975, AY922324–AY922344, EF197846–EF197847, and EF203471.

Acknowledgments

BAC libraries were generated in collaboration with Thomas Mitchell-Olds, Heiko Vogel, and Jürgen Kroymann (Max Planck Institute for Chemical Ecology, Jena). We thank Stefan Schuster for advice regarding BAC sequencing, Marcus Koch for advice regarding genus affiliations and phylogeny of the Brassicaceae species, Sang-Tae Kim for help with phylogenetic analyses, Javier Palatnik for discussion and the MIR319c overexpression construct, Heike Wollmann for sharing technical expertise, and Javier Palatnik, Heike Wollmann, and especially Juliette de Meaux for discussions and critical reading of the manuscript. Work on small RNAs in the Weigel laboratory is...
supported by European Community FP6 IP SIROCCO (contract LSHG-CT-2006-037900) and by the Max Planck Society, of which D.W. is a Director.

Literature Cited


Swofford DL. 1993. PAUP: a computer program for phylogenetic inference using maximum parsimony. J Gen Physiol. 102:9A.


Peter Lockhart, Associate Editor
Accepted January 30, 2008