Structurally different alleles of the ath-*MIR824* microRNA precursor are maintained at high frequency in *Arabidopsis thaliana*

Juliette de Meaux†, Jin-Yong Hu, Ute Tartler, and Ulrike Goebel

Max Planck Institute for Plant Breeding Research, Carl-von-Linné Weg 10, 50829 Cologne, Germany

Communicated by Maarten Koornneef, Wageningen University and Research Centre, Wageningen, The Netherlands, April 2, 2008 (received for review December 21, 2007)

In plants and animals, gene expression can be down-regulated at the posttranscriptional level by microRNAs (miRNAs), a class of small endogenous RNA. Comparative analysis of miRNA content across species indicates continuous birth and death of these loci in the course of evolution. However, little is known about the microevolutionary dynamics of these genetic elements, especially in plants. In this article we examine polymorphism at two miRNA-encoding loci in *Arabidopsis thaliana*, miR856 and miR824, which are not found in rice or poplar. We compare their diversity to other miRNA-encoding loci conserved across distant taxa. We find that levels of variation vary significantly across loci and that the two recently derived loci harbor patterns of diversity deviating from neutrality. miRNA miR856 shows a weak signature of a selective sweep whereas miR824 displays signs of balancing selection. A detailed examination of structural variation among alleles found at the miR824-encoding locus suggests nonrandom evolution of a thermoresistant substructure in the precursor. Expression analysis of pre-miR824 and its target, AGL16, indicates that these structural differences likely impact the processing of mature miR824. Our work highlights the relevance of RNA structure in precursor sequence evolution, suggesting that the evolutionary dynamics of miRNA-encoding loci is more complex than suggested by the constraints exerted on the interaction between mature miRNA fragments and their target exon.

balancing selection | microRNA evolution | secondary structure

MicroRNAs (miRNAs) are endogenous small RNA that negatively regulate gene expression in plants and animals. These 20- to 24-nt RNAs are processed from hairpin-forming primary transcripts (pre-miRNAs) and direct the down-regulation of specific mRNA targets.

The comparative analysis of miRNA contents in diverse animal or plant species has revealed both long-term maintenance and taxa-specific occurrence of miRNA genes, indicating a birth-and-death evolutionary dynamics (1, 2). In *Arabidopsis thaliana*, 21 miRNA families are also found in rice, and miRNA/target pairing can be maintained over evolutionary time (3, 4, 6). Now, thanks to high-throughput small RNA sequencing, taxa-specific small RNAs are about to outnumber so-called conserved miRNAs (1, 6–9). The dynamic evolution of miRNA-encoding genes is further supported by the detailed study of miR319 variation in the Brassicacae (10). A similar picture emerges from the comparative analysis of small RNAs in humans and chimpanzees (11, 12).

Numerous differences in both biogenesis and mode of miRNA action were found between plants and animals, suggesting that small-RNA regulatory systems have evolved largely independently in the two kingdoms (see ref. 13 for a recent review). Interestingly, these mechanistic differences may cast different evolutionary constraints on the evolution of miRNAs in plants and animals. Studying the dynamics of emergence of new miRNAs offers a unique chance to elucidate the evolutionary commonalities and differences that characterize the history of two regulatory systems evolving in parallel (14).

It has been proposed that new miRNAs could evolve by point mutation in animals, followed by selection against inadequate miRNA/mRNA pairing (14). Such a scenario is highly unlikely in plants because miRNA/target pairing extends over twice as many nucleotides. Instead, evolution by inverted duplication has been proposed (15). Especially intriguing in this model is the idea of a step-by-step evolution. It implies that newly formed expressed RNA hairpins may experience successive selective sweeps until they reach an evolutionarily optimized form, allowing a one-to-one miRNA/target relationship to be established. In both animals and plants, recently derived miRNAs are typically expressed at low levels, which may help maintain negative pleiotropic effects at low levels until compensatory mutations have emerged (14).

Therefore, it seems that the evolution of new miRNAs might be dominated by evolution in the pre-miRNA-encoding region in plants. By contrast, in animals, new miRNAs require selection against pairing in unwanted target mRNAs and expression modifications (14). At this point, however, scenarios for miRNA evolution can only be tentative because existing information on miRNA evolutionary dynamics is scanty. To the best of our knowledge, there are no data on polymorphism segregating at recently derived miRNA loci in plants.

The examination of structural evolution in novel miRNAs promises to be particularly insightful because the relationship between sequence and structure can be examined with exquisite accuracy at these loci (16). Indeed, miRNAs are processed from an expressed single-strand RNA folded into a hairpin (the pre-miRNA fold-back), and the physical parameters controlling RNA folding are exceptionally well understood. Therefore, a glimpse into the fitness landscape associated with sequence variation is possible. Thanks to this property, >80% of metazoan miRNA precursors could be shown to have evolved significant robustness in their hairpin structure (17). Second, RNA folding dynamics are highly temperature-dependent, and pre-miRNA hairpin formation may be influenced by temperature. In plants, this capacity may provide a mechanism for expressing temperature-specific plasticity. Conversely, the evolution of temperature-independent pre-miRNA structures may be especially important in those species. Surprisingly, this matter has been the subject of hardly any attention so far.

In this article we ask the following questions: (*i*) Do recently derived miRNA-encoding loci show singular evolutionary dynamics? (*ii*) Is there evidence for evolutionary optimization of young

Author contributions: J.d.M. and U.G. designed research; J.d.M., J.-Y.H., U.T., and U.G. performed research; J.d.M., J.-Y.H., and U.G. analyzed data; and J.d.M. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. FM163680–FM164275).

†To whom correspondence should be addressed. E-mail: demeaux@mpiz-koeln.mpg.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0803218105/DCSupplemental.

Table 1. Summary of polymorphisms found in pre-miRNA loci

Locus	Length, bp	π	θ	K	Divergence in matured miRNA	Polymorphism in matured miRNA[†]
pre-miR164b	171	0.003	0.003	0.067	0	1 (f=0.04)
pre-miR167a	195	0.002	0.005	0.056	0	0
pre-miR167b	109	0.000	0.000	0.009	0	0
pre-miR171b	117	0.001	0.006	0.018	0	0
pre-miR172a	102	0.000	0.000	0.020	0	0
pre-miR172b	95	0.000	0.000	0.055	1	0
pre-miR319a	176	0.001	0.003	0.012	0	0
pre-miR390a	107	0.000	0.000	0.058	0	0
pre-miR393a	133	0.012	0.009	0.072	0	1 (f=0.02)
pre-miR395a	93	0.001	0.005	-	—	0
pre-miR395b	100	0.001	0.007	0.021	—	0
pre-mir395e	95	0.010	0.008	0.052	0	0
pre-mir395f	116	0.001	0.002	0.056	0	1 (f=0.02)
pre-miR396b	135	0.003	0.002	-	—	0
pre-miR824	734	0.026	0.015	0.056	0	0
pre-miR856	272	0.002	0.004	0.080	2	0

[†]Frequency is given for each derived polymorphism found in the matured pre-miRNA.

pre-miRNA stem–loop structures? (*iii*) Can pre-miRNA structural variation impact miRNA function?

To this aim, we sequenced two loci encoding recently derived miRNAs [miR824 and miR856 (1)] in ≈40 *A. thaliana* ecotypes and compared this diversity to that found at 14 miRNA loci conserved in rice, poplar, and *A. thaliana*. We analyzed multilocus patterns of polymorphism and divergence. Our work demonstrates that pervasive variation occurs at miRNA-encoding loci in *A. thaliana*, yet this variation is not uniformly distributed across loci. In our study, the two recently derived miRNA loci display evolutionary histories that differ significantly from that observed at 14 ancient miRNAs. We found two haplotypes segregating at high frequency at locus miR824. Although these haplotypes harbor identical 21-bp miRNA fragments, we were able to show that one allele has evolved a temperature-resistant secondary structure. Studies of target gene expression indicate that differences in the pre-miRNA secondary structure affect the efficiency of target down-regulation. Our study shows that selection not only acts on the mature miRNA fragment but can also considerably influence the evolution of foldback structure in the pre-miRNA.

Results

We obtained information on diversity segregating in *A. thaliana* at 16 loci encoding miRNA precursors (pre-miRNA), two of which are lineage-specific. miRNA856 is so far known only in *A. thaliana* and *Arabidopsis lyrata*, and miR824 was found only in the *Brassicacae* (18). The length of pre-miRNA ranged from 95 to 734 bp. We observed 0–36 segregating sites and diversity, as measured by the average number of pairwise differences (π), ranged from 0.000 to 0.026 (Table 1).

Variability in Evolutionary History Across Loci. We conducted a population genetics analysis of variation to evaluate the recent evolutionary history of the different miRNA loci. The maximum likelihood ratio method developed by Schmid *et al.* (19) was used to estimate variability in population mutation rate across the 16 pre-miRNA loci of this study [supporting information (SI) Table S1]. The best model ($P = 0.026$) suggests the existence of two classes of miRNAs, a major one with a low mutation rate ($\theta = 0.003$) and a smaller class with a relatively high mutation rate ($\theta = 0.0138$). The latter class contained pre-miR393a, pre-miR395e, and pre-miR824 (Table 1). To examine whether this pattern is due to differences in mutation rate across loci, we performed a multilocus Hudson–Kreitman–Aguadé (HKA) test comparing the ratio of polymorphism to divergence across loci (20). This test was significant ($\chi^2 = 48.9286$, df = 13, $P < 0.001$), indicating that the evolutionary history varied across the 14 loci for which an outgroup sequence could be identified (Table S2). Measures of partial HKA reveal the part of multilocus evolutionary heterogeneity attributable to each single locus and showed that pre-miR824 and pre-miR856 show evolutionary histories that depart from the remaining loci (Table S2, 59% and 9.2% of the total HKA χ^2, respectively). This result was obtained while considering the whole region within and around predicted pre-miRNAs. We conducted the same analysis on only pre-miRNA sites, and no significant multilocus HKA was observed ($\chi^2 = 16.54$, df = 13, $P > 0.2$). However, partial HKA for miR824 remained high, indicating that this result may reflect lowered power when considering only part of the data ($\chi^2 = 4.588$, 27% of total HKA). A significant multilocus HKA was detected in flanking regions (HKA $\chi^2 = 42.5$, df = 24, $P = 0.01$; Table S3). The analysis of partial HKA indicated that the upstream region of miR856 made 39.8% of the total HKA value (data not shown).

pre-miR824 and pre-miR856 appear to have different evolutionary histories. pre-miR856 shows relatively low levels of polymorphism but the highest level of divergence, whereas pre-miR824 has the highest level of polymorphism and the lowest level of divergence (Table 1 and Table S4). Consistent with the result of the multilocus HKA test, the detailed analysis of summary statistics and neutrality tests conducted over all pre-miRNAs and flanking regions revealed that the two loci are consistent outliers (Table 2 and Table S5).

First pre-miR856 showed the most negative Tajima's D value across pre-miRNA-encoding loci examined, especially when both pre-miRNA and flanking regions were analyzed together ($D = -2.039$). Negative values of Tajima's D indicate an excess of low-frequency mutations and can be an indication of a recent selective sweep. This value, however, falls within the confidence interval calculated for Tajima's D determined by multilocus analysis at 56 noncoding loci in *A. thaliana* [average $D = -1.035$, 95% confidence interval (−2.528; 0.458) recalculated from the data by Nordborg *et al.* (21) on our sample of ecotypes]. This locus also shows the most significant Fay and Wu H value, indicating an excess of high-frequency-derived mutation ($H = -16.018$, $P < 0.001$). This pattern results from a single very divergent allele observed in accession Can-0 and harboring the ancestral state at multiple polymorphic sites. The distribution of Fay and Wu's H in *A. thaliana* appears to be relatively unaffected by recent demographic events in this species (19).

Patterns of diversity at pre-miR824 showed a clearer departure from neutral expectations. First, Tajima's D was high at this locus ($D = 2.11$). This value is significantly outside of the confidence interval obtained in the multilocus analysis [average $D = -1.035$, 95% confidence interval (−2.528; 0.458) recalculated from the data

EVOLUTION

Table 2. Summary statistics and tests of neutrality performed on pre-miRNA coding genes and their flanking regions

Locus	No. of sequences	No. of sites	Flanking region, bp	*K*	*H*	*F*	*D*
pre-miR164b	43	265	113	0.066	0.300	−0.798	−0.202
pre-miR167a	45	248	60	0.064	−3.421**	0.431	−1.309
pre-miR167b	34	967	912	0.062	−1.825	−1.180	−1.229
pre-miR171b	47	365	306	0.082	−1.505*	−1.799	−1.655**
pre-miR172a	38	401	311	0.060	0.144	−1.145	−1.296
pre-miR172b	42	341	277	0.059			
pre-miR319a	19	3005	3111	0.113	−2.30994	−0.926	−0.877
pre-miR390a	47	269	164	0.068			
pre-mir393a	14	920	787	0.119	0.078078	−0.530	−0.465
pre-miR395a	45	478	408				−1.333
pre-miR395b	42	474	381	0.088	−0.999	−2.338*	−1.484*
pre-miR395e	31	421	351	0.094	1.024	−0.178	−0.068
pre-miR395f	31	532	443	0.107	−1.55269	−1.553	−0.310
pre-miR396b	41	773	643				0.497
pre-miR824	46	633	168	0.053	1.244	0.699	2.115(*)
pre-miR856	40	457	204	0.108	−16.018***	−1.945*	−2.039**

H, *F*, and *D* are three neutrality tests characterizing the frequency spectrum in the population (see *SI Text*). *, $P = 0.05$; **, $P = 0.01$; ***, $P = 0.001$. Significance was calculated by coalescent simulations, except if indicated between parentheses, where the statistics are significantly different from the empirical distribution drawn by multilocus analysis (see *Materials and Methods*). *K* is the divergence rate between *A. thaliana* and *A. lyrata*.

by Nordborg *et al.* (21) on our sample of ecotypes]. In addition, it falls within the 2% highest values of Tajima's *D* obtained on the same sample of accessions at 864 coding and noncoding loci (21). This pattern is due to the segregation of two major allelic clades, ≈711 and 824 nt long, hereafter mentioned as short-allele and long-allele clades, respectively. The two clades differ by 27 fixed nucleotide differences and six fixed indels, one of which is 103 bp long. Each clade is represented by equivalent numbers of accessions (Fig. 1). Interestingly, the short-allele clade shows a strongly negative Tajima's *D* value ($D = -1.826$), indicating an excess of low-frequency-derived mutations, whereas the long-allele clade is almost monomorphic. A sliding-window analysis shows that most of the divergence has occurred in the loop region of the pre-miRNA (Fig. 2). Yet each clade shows higher divergence in different parts of the loop. The long-allele clade appears to have accumulated divergence mutations in the 3′ part of the loop, which is deleted in the short-allele clade. Instead, in the short-allele clade mutations have accumulated in the 5′ part of the loop (Fig. 2). No variation was observed in either the miRNA or the miRNA-complementary fragments (miRNA*).

Structural Variation at Locus miR824. We investigated variation in hairpin structure between the two clades. Structural predictions are temperature-dependent; therefore, we examined RNA folding at 3°C, 20°C, and 37°C to model the temperature range that an individual plant is likely to meet along the season. Secondary structure can be represented on a matrix indicating whether or not two nucleotides are paired in the structure presenting Gibbs' minimum free energy (MFE) (Fig. 3, below diagonal in each diagram). However, Gibbs' MFE provides only a raw prediction of secondary structure because it does not take into account the presence of alternative structures competing with the most stable structure (16). The matrix of base pair probabilities in thermodynamic equilibrium (Fig. 3, above diagonal) gives a more complete picture of the range of structures that is accessible to a sequence (22). Several authors have derived a measure of well definedness from this matrix, that is, of the relative abundance of a single structure or substructure in the ensemble of all possible structures (23). The established position-wise measures of this kind do not take into account whether or not a well defined substructure is part of the MFE structure (24). We developed a measure reflecting both the stability and probability of occurring readily of a structure, which can be applied to segments of the whole structure. This measure is the Pearson correlation coefficient between Gibbs' MFE structure and measures of well definedness (see *Materials and Methods*). If the well defined structure is in agreement with the MFE structure the coefficient is close to 1 whereas values close to 0 reflect poorly defined structures, for which the MFE structure delivers little information.

Fig. 3 shows that for each clade both branches of the stem have a high probability of pairing associated with low free energy.

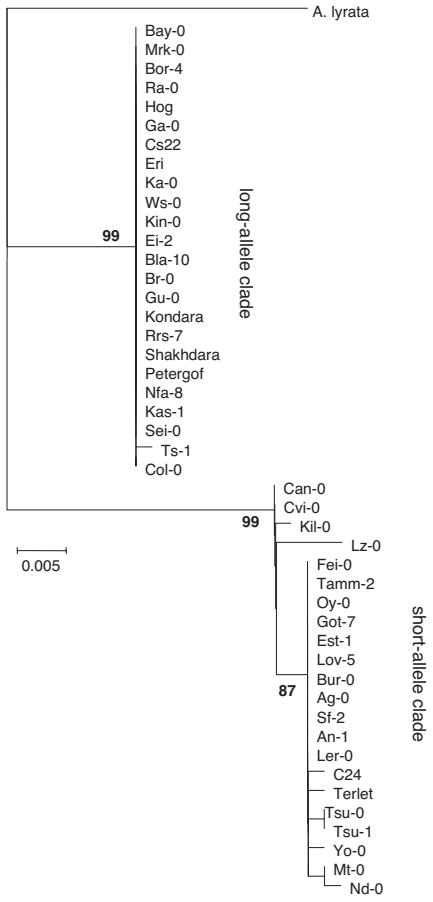

Fig. 1. Neighbor-joining tree showing diversity at the MIR824 locus in *A. thaliana*. Bootstrap values were calculated with 5,000 permutations. Note that no recombination was detected in this sample.

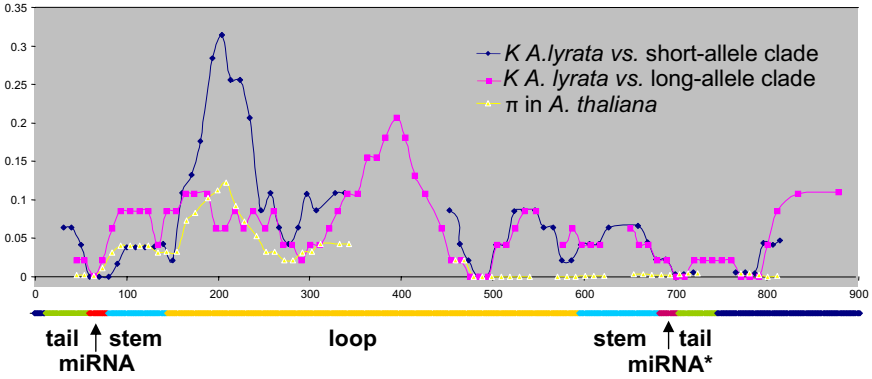

Fig. 2. Sliding-window analysis of polymorphism and divergence at locus pre-miR824, within and between the allelic clusters in *A. thaliana* and the orthologous sequence in *A. lyrata*. Areas where indels occur are revealed by discontinuities in the curve.

Therefore, the two allelic types are equally likely to be functional, despite elevated levels of divergence. Indeed, mature miR824 was detected in both Col-0 and Ler-0 ecotypes, which harbor the long and short allele, respectively (18). In the short allele, a secondary stem–loop structure appears, ≈100 bp long, that is well defined (high probability of pairing) and part of the MFE structure (Fig. 3, "Box 2" structure, and Fig. S1). This structure is totally absent in the long alleles at all temperatures as illustrated by the fuzzy pairing probability in the corresponding area (Fig. 3, long allele, discontinuous arrow). Instead, a well defined secondary structure that is part of the MFE structure forms in another part of the sequence (Fig. 3, "Box 3" structure). This structure is absent from the other clade (Fig. 3, short allele, discontinuous arrow); it is different from the "Box 2" structure and harbors several branches (Fig. S2). The definedness and stability of these structures in each allelic clade are quantified by Pearson correlation coefficients and are reported in Table 3. These coefficients provide a good summary of the structural differences between the two allelic clades, with the "Box 2" motif being thermoresistant only in the short-allele clade and the "Box 3" structure being most likely to exist at low temperatures only in the long-allele clade (Table 3).

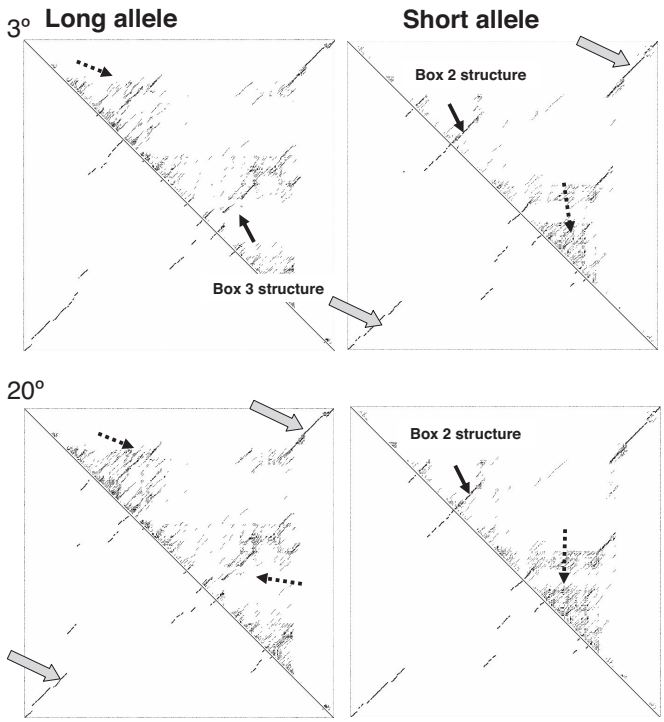

Fig. 3. Dot plots illustrating structural definedness for two alleles (Col-0 and C24) representative of the long- and short-allele clades at two temperatures, respectively. Discontinuous arrows show poorly defined secondary structures with weak correlation coefficients between the probability of pairing and the MFE structure. Continuous arrows show well defined and stable secondary structures with highly correlated probabilities of pairing and actual pairing in the MFE structure. Block arrows show that areas around miRNA and miRNA* form a well defined and stable hairpin in both alleles.

Regions of Homology Between miR824 and Its Gene of Origin Localize to Structurally Important Elements. miRNA miR824 shows extended homology with its target gene, AGL16, from which it presumably originated (1). The full length of the pre-miRNA sequence aligns with the 5′ part of the reverse-complemented AGL16 genomic sequence. A partial duplication of AGL16 could have been sufficient to create miR824 because the two regions homologous to miRNA and miRNA* in AGL16 readily pair together (Fig. S3). Five boxes can be highlighted that are conserved between AGL16 and pre-miR824, two of which are involved in the substructures described above (Fig. S4).

Intriguingly, the region that pairs with box 2 in the short allele, to form the above-mentioned "Box 2" structure, corresponds to the peak of divergence between the *A. lyrata* allele and the short-allele clade (Fig. 2), whereas the divergence peak between *A. lyrata* and the long-allele clade falls in the region that pairs with box 3, forming the above-mentioned "Box 3" structure. Therefore, regions of accelerated divergence in each allelic clade map at well defined and most stable secondary structures and pair with conserved sequence elements. Analysis of structural definedness and stability in *A. lyrata* and in the putative ancestral sequence suggests that the thermoresistant "Box 2" structure and the absence of the "Box 3" structure are derived in the *A. thaliana* short-allele clade (Table 3).

Of all mutations differentiating the two clades, only a few mutations are sufficient to promote the formation of the "Box 2" structure. These mutations are (*i*) the insertion of GCT in the strand of the substructure that is highly conserved with AGL16 and (*ii*) three of five closely spaced mutations on the partner strand of box 2, which help to accommodate the GCT insertion into a longer stem (Fig. S1). The GCT insertion seems to be the result of a replication slippage because it occurred in an existing TGC sequence. The cluster of compensatory mutations on the opposite strand is consistent with the hypothesis of selection for the formation of the structural element.

Spatially Heterogeneous Distribution of Variation at miR824. We investigated an effect of local temperature on the distribution of polymorphism at pre-miR824. First, we used our population genetics sample to test association between allelic polymorphism and environmental climatic variable at the location of origin of each of 41 accessions. We found that variation in the first principal component of average climatic variables can be explained by the allelic clades ($F_{1,40} = 8.507$, $P = 0.006$). The first component mostly represents temperature variables (see *SI Methods*).

EVOLUTION

Table 3. Pearson correlation coefficient measuring the concomitant definedness and stability of secondary structures in the pre-miR824 sequences in *A. thaliana* (C24 and Col-0) and *A. lyrata* at different temperatures (described in *Materials and Methods*)

		Pearson R^2			
Structure	Temperature, °C	*A. thaliana* short allele (C24)	*A. thaliana* long allele (Col-0)	*A. lyrata*	*A. thaliana* (ancestral)
Box 2	3	0.87	0.00	0.83	0.25
	20	0.91	0.09	0.27	0.66
	37	0.88	0.46	0.80	0.82
Box 3	3	0.42	0.95	0.53	0.42
	20	0.53	0.49	0.96	0.62
	37	0.60	0.81	0.47	0.56

C24 and Col-0 are representative of the short- and long-allele clades, respectively. The ancestral sequence was reconstructed as described in *Materials and Methods*.

To confirm this effect on a larger sample of accessions, we genotyped intermediate-frequency SNPs at these loci on 192 *A. thaliana* accessions collected throughout the species range (see *SI Methods*). To our surprise, no significant association could be found ($F_{1,136} = 0.47$, $P > 0.5$). However, this sample contains a large proportion of accessions from the central part of the species range. Association in the previous sample may be due to a more uniform distribution of samples throughout the latitudinal range. We subsequently genotyped miR824 alleles within seven populations collected in Spain as well as within 13 populations collected throughout Norway to compare diversity between the two extremes of the species' range. Of these 13 populations, there were two polymorphic populations, the 11 remaining populations being fixed for the short allele. By contrast, in Spain, only one population was fixed for the short allele, all other six populations being polymorphic or fixed for the long allele.

The Long pre-miR824 Allele Is Less Efficient at Down-Regulating AGL16. To assess functional differences between the long and short alleles at locus miR824, we used two lines, line 62/8/8 and Col-0, which are identical on all chromosomes except for 800 kb in the region of miR824 on chromosome 4. These two lines are therefore identical at the locus encoding AGL16, the target of miR824. We analyzed expression levels of full-length AGL16 as well as full-length pre-miR824 in leaves of plants grown at 20°C. Full-length AGL16 expression was lower in line 62/8/8 harboring the short allele at locus miR824 (two-tailed *t* test; $t = 5.877$, $P = 0.004$). Because expression levels of pre-miR824 are higher in this line than in Col-0 (two-tailed *t* test; $t = 8.00$, $P = 0.001$), increased miR824-mediated AGL16 cleavage in line 62/8/8 can only be due to increased efficiency of miR824 processing in the short allele. Indeed, both alleles harbor identical mature miR824 fragments.

Discussion

Evolution of Recently Derived miRNA Loci. Our analysis of multiple miRNA-encoding loci in *A. thaliana* reveals that the two recently derived miRNA loci have singular evolutionary trajectories (Table 2 and Table S5).

At pre-miR856, the departure from neutrality is marginal. However, it is interesting to note that miR856 shows the highest level of divergence ($K = 0.08$ and $\pi = 0.002$) despite a low level of polymorphism (Table 1), and its upstream flanking region causes a significant multilocus HKA across the flanking regions examined in this study (Table S3). This locus harbors the largest number of fixed nucleotide changes in the mature miRNA fragment. This cannot be explained by a larger mutation rate because the miR856 locus belongs to low-diversity pre-miRNA loci. CHX18 is a likely target for miR856 (1). The target site is identical in both *A. thaliana* and *A. lyrata* (data not shown). In both species, each mature miR856 fragment has one mismatch to the target, at positions 7 and 5 of the miRNA, respectively. Because one mismatch in the 5′ region is not expected to compromise cleavage (25), the variation between the two species at miR856 suggests the occurrence of compensatory mutations.

At locus miR824 instead, patterns of diversity depart unambiguously from neutral expectations (Table 2, Table S5, and Fig. 1). Two clades with very divergent alleles of different length are found to segregate at intermediate frequencies (Fig. 1). This pattern suggests that balancing selection may maintain these two allelic types at high frequency in the population.

Clearly, patterns of pre-miR824 structural variation within and between allelic clades in *A. thaliana* are not random and warrant further investigation. The short-allele clade harbors a secondary stem that is strikingly well defined across a large temperature range ("Box 2" structure, Fig. 3, Table 3, and Fig. S1). One strand of the stem corresponds to the peak of divergence with the *A. lyrata* allele whereas the second strand shows low divergence (Fig. 2).

In the case described here, it is especially interesting to note that this structure is absent in the long-allele clade. This latter instead harbors a distinct well defined secondary structure ("Box 3" structure, Fig. 3 and Fig. S2). Again, one strand of the structure corresponds to the peak of divergence between this allele and *A. lyrata*. The exclusive formation of each of these secondary substructures in each of the two intermediate-frequency clades is consistent with the hypothesis that each of these secondary stems may provide different selective advantages, a pattern expected under balancing selection. Indeed, our expression studies in two isogenic lines harboring different alleles at the miR824-encoding locus show that levels of expression of the target gene are not lower when the long-allele pre-miR824 is expressed at higher levels than the short allele, indicating that substructural differences impact how the mature miR824 fragment is processed. We further observed that the "Box 2" secondary structure was reminiscent of a hairpin (Fig. S1). We indeed found that a 21-bp fragment in this structure has a potential target in gene AT3G45630 (see *SI Methods*). Precursors cleaved into multiple independent small RNAs have been described in diverse taxa (26, 27), yet we were not able to confirm the existence of a cleaved AT3G45630 mRNA in leaf tissue of ecotypes carrying the short allele at the miR824 locus.

There were two reasons to investigate whether polymorphism at pre-miR824 may reflect climatic variation across the range of *A. thaliana*. First, miR824 controls stomata density, and demands on this trait may vary across the climatic range of the species (18). Second, the alleles in the long-allele clade display a secondary structure that is best-defined and most stable at 3°C (see Table 3). We could not find a consistent effect of average climatic parameters on variation at miR824 in *A. thaliana*, yet we find that miR824 variation is not geographically uniform. The relative roles played by local selection and population structure in this pattern will probably be hard to disentangle. The pattern of geographical distribution of the polymorphism at pre-miR824 is complex, with populations in Norway being predominantly fixed for the short allele and populations in Spain being mostly polymorphic. In any case, this suggests that the two alleles can compete locally, and further investigation is

needed to examine whether year-to-year climatic variance may explain balanced levels of polymorphism.

Conclusion

Our work demonstrates that pervasive variation occurs at miRNA-encoding loci and illustrates how the detailed study of the hairpin backbone can reveal intriguing patterns of nonrandom variation. In the small set of miRNA-encoding loci included in this study, the two recently derived miRNAs show deviating patterns of diversity. As this article was in revision, polymorphism data were reported for 66 loci encoding miRNA conserved across distant taxa (28). Although conducted on a smaller sample of *A. thaliana* accessions, this study yielded essentially the same levels of diversity as observed in our study and confirms that the evolutionary dynamics of miR856 and miR824 differs from that observed for miRNAs conserved among distant taxa. Although our results clearly point to an adaptive dynamic specific to recently derived miRNA, as was also recently reported in *Drosophila* (29), it is too early to elaborate a general scenario. Indeed, patterns of diversity indicate different evolutionary regimes for the two recently derived genes. Although both are young with respect to the other loci examined in this and the other study (28), these two miRNA genes may have different ages. miR824 was shown to target AGL16 in *Brassica rapa* and *Brassica napus* and probably arose some 24 millions years ago (18, 30). miR856 is also not found in *Populus* or rice, but it is not known whether it is conserved in the Brassicacae or younger. Upcoming genomic data in the Brassicacae will allow examining in detail the adaptive role of a larger number of recently derived miRNAs. We believe that miRNA-encoding loci should not be ignored in the endeavor to identify the molecular basis of genetic adaptation.

Materials and Methods

Population Genetic Analyses. Sequences were obtained as described in *SI Methods* and Tables S6 and S7 and aligned with Megalign 5.03 (DNASTAR). The DnaSP 4.0 program (31) was used for both intraspecific and interspecific analyses of nucleotide polymorphism. Multilocus analyses were performed by using the software MANVa (www.ub.edu/softevol/manva/), which implements the standardized Fay and Wu *H* test, measuring an excess of high-frequency mutations in the population (32). MANVa also implements a maximum-likelihood estimation of mutation rates across loci (see ref. 19 for a detailed description). A multilocus HKA test is implemented in MANVa to compare the ratio of intraspecific polymorphism to interspecific divergence across multiple loci. The software further calculates partial HKA, the discrepancy between observed and expected value, for each locus to identify the loci responsible for heterogeneity in evolutionary rates. HKA tests were performed for all positions in the pre-miRNA and flanking regions. An experimental distribution of Tajima's *D* expected in *A. thaliana* at noncoding loci was drawn from the analysis of 56 intergenic loci sequenced by Nordborg *et al.* (21). To correct for possible sampling differences, we conducted the multilocus analysis on the subsample of accessions used in this study. A detailed description of all summary statistics is given in Table S5. A putative allele ancestral to the two clades observed at locus miR824 in *A. thaliana* was reconstructed by parsimony and harbored the *A. lyrata* sequence at deleted positions.

Analysis of Structural Evolution in pre-miR824. We examined variation in the formation of well defined substructures in the miR824 pre-miRNA foldbacks using *in silico* folding predictions, which we performed with the RNAfold.pl program, a Perl interface to the C folding functions delivered with ViennaRNA-1.6 (16). Structural predictions were performed by using standard parameters and varying only the folding temperature. Structural predictions are displayed on dot plots where sequence positions run from left to right and from top to bottom. The ensemble base pair probability is shown above the diagonal, with black squares of size proportional to the probability, and the nucleotide pairs found in the MFE structure are represented below the diagonal. Well defined and stable substructures are defined as those in which bases that pair in the MFE secondary structure have a high probability of pairing in the ensemble of possible structures (see *SI Methods*). We quantified this by the Pearson correlation coefficient between the ensemble base pair probability and the MFE structure in the region of the structure considered.

Gene Expression Analysis in Near-Isogenic Lines. Near-isogenic lines derived from crosses between Col and C24 were provided by T. Altmann (Max Planck Institute of Molecular Plant Physiology, Golm, Germany). Line 62/8/8 contains a C24 introgression around markers IV81 and IV82 in the Col-0 background and contains a short allele at the miR824-encoding locus (5). Plants were grown under long-day conditions (16 h light, 20°C/8 h dark, 16°C). Gene expression levels were analyzed by semiquantitative RT-PCR as described in *SI Methods*.

ACKNOWLEDGMENTS. We thank Sebastian Ramos-Onsins for advice in the multilocus analysis of population variation, Carlos Alonso-Blanco and the Norwegian *Arabidopsis* Research Center for providing seeds from local *A. thaliana* populations in Spain and Norway, the Department of Energy's Joint Genome Institute for shotgun data from the *A. lyrata* genome sequencing effort, and Marilyne Debieu for climatic information at the location of origin of the accessions used in this study.

1. Fahlgren N (2007) High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of MIRNA genes. *PLoS ONE* 2:e219.
2. Lu J, *et al.* (2008) The birth and death of microRNA genes in Drosophila. *Nat Genet* 40:351–355.
3. Floyd SK, Bowman JL (2004) Gene regulation: Ancient microRNA target sequences in plants. *Nature* 428:485–486.
4. Jones-Rhoades MW, Bartel DP (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell* 14:787–799.
5. Törjék O, *et al.* (2003) Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis. Plant J* 36:122–140.
6. Rajagopalan R, Vaucheret H, Trejo J, Bartel DP (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev* 20:3407–3425.
7. Fattash I, Voss B, Reski R, Hess WR, Frank W (2007) Evidence for the rapid expansion of microRNA-mediated regulation in early land plant evolution. *BMC Plant Biol* 7:13.
8. Axtell MJ, Bartel DP (2005) Antiquity of microRNAs and their targets in land plants. *Plant Cell* 17:1658–1673.
9. Barakat A, Wall PK, Diloreto S, Depamphilis CW, Carlson JE (2007) Conservation and divergence of microRNAs in *Populus. BMC Genomics* 8:481.
10. Warthmann N, Das S, Lanz C, Weigel D (2008) Comparative analysis of the MIR319a microRNA locus in Arabidopsis and related Brassicacae. *Mol Biol Evol* 25:892–902.
11. Bentwich I, *et al.* (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet* 37:766–770.
12. Berezikov E, *et al.* (2006) Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* 38:1375–1377.
13. Chapman EJ, Carrington JC (2007) Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet* 8:884–896.
14. Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* 8:93–103.
15. Allen E, *et al.* (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nat Genet* 36:1282–1290.
16. Hofacker IL, *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte Chem* 125:167–188.
17. Borenstein E, Ruppin E (2006) Direct evolution of genetic robustness in microRNA. *Proc Natl Acad Sci USA* 103:6593–6598.
18. Kutter C, *et al.* (2007) MicroRNA-mediated regulation of stomatal development in Arabidopsis. *Plant Cell* 19:2417–2429.
19. Schmid KJ, *et al.* (2005) A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* 169:1601–1615.
20. Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.
21. Nordborg M, *et al.* (2005) The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol* 3:1289–1299.
22. McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29:1105–1119.
23. Hofacker IL (2005) *RNA Secondary Structure Prediction* (Wiley, Hoboken, NJ).
24. Huynen MA, Perelson A, Vieira W, Stadler P (1996) Base-pairing probabilities in a complete HIV-1 genome. *J Comput Biol* 3:253–274.
25. Schwab R, *et al.* (2005) Specific effects of microRNAs on the plant transcriptome. *Dev Cell* 8:517–527.
26. Lu SF, Sun YH, Amerson H, Chiang VL (2007) MicroRNAs in loblolly pine (Pinus taeda L.) and their association with fusiform rust gall development. *Plant J* 51:1077–1098.
27. Stark A, *et al.* (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* 450:219–232.
28. Ehrenreich IM, Purugganan MD (2008) Sequence variation of microRNAs and their binding sites in *Arabidopsis thaliana. Plant Physiol* 146:1974–1982.
29. Lu J, *et al.* (2008) Adaptive evolution of newly emerged micro-RNA genes in Drosophila. *Mol Biol Evol* 25:929–938.
30. Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). *Mol Biol Evol* 17:1483–1498.
31. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.
32. Zeng K, Fu YX, Shi S, Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174:1431–1439.