

Conservation and divergence of plant microRNA genes

Baohong Zhang¹, Xiaoping Pan¹, Charles H. Cannon², George P. Cobb¹ and Todd A. Anderson^{1,*}

¹The Institute of Environmental and Human Health (TIEHH), and Department of Environmental Toxicology, Texas Tech University, Lubbock, TX 79409-1163, USA, and

²Department of Biological Science, Texas Tech University, Lubbock, TX 79409-3131, USA

Received 1 December 2005; accepted 20 December 2005.

*For correspondence (fax +1 806 885 4577; e-mail todd.anderson@ttu.edu).

Summary

MicroRNA (miRNA) is one class of newly identified, small, non-coding RNAs that play versatile and important roles in post-transcriptional gene regulation. All miRNAs have similar secondary hairpin structures; many of these are evolutionarily conserved. This suggests a powerful approach to predict the existence of new miRNA orthologs or homologs in other species. We developed a comprehensive strategy to identify new miRNA homologs by mining the repository of available ESTs. A total of 481 miRNAs, belonging to 37 miRNA families in 71 different plant species, were identified from more than 6 million EST sequences in plants. The potential targets of the EST-predicted miRNAs were also elucidated from the EST and protein databases, providing additional evidence for the real existence of these miRNAs in the given plant species. Some plant miRNAs were physically clustered together, suggesting that these miRNAs have similar gene expression patterns and are transcribed together as a polycistron, as observed among animal miRNAs. The uracil nucleotide is dominant in the first position of 5' mature miRNAs. Our results indicate that many miRNA families are evolutionarily conserved across all major lineages of plants, including mosses, gymnosperms, monocots and eudicots. Additionally, the number of miRNAs discovered was directly related to the number of available ESTs and not to evolutionary relatedness to *Arabidopsis thaliana*, indicating that miRNAs are conserved and little phylogenetic signal exists in the presence or absence of these miRNAs. Regulation of gene expression by miRNAs appears to have existed at the earliest stages of plant evolution and has been tightly constrained (functionally) for more than 425 million years.

Keywords: cluster, EST, evolution, microRNA, origin, plant.

Introduction

MicroRNAs (miRNAs) are a newly identified class of non-coding, approximately 21–23 nucleotide-long, small RNAs (Ambros, 2001), originating from long self-complementary (foldback) precursors (pri-miRNAs) (Bartel, 2004). Mature miRNA formation is a multi-step process involving many complicated enzymes (Bartel, 2004). First, pri-miRNAs are cleaved to miRNA precursors with a characteristic hairpin structure. This step is catalyzed by the RNaseIII-like endonuclease Droscha and Dicer in animals (Lee *et al.*, 2003) or by a Dicer-like enzyme (DCL) in plants (Park *et al.*, 2002; Reinhart *et al.*, 2002). Then, pre-miRNA is further cleaved to a miRNA duplex (miRNA:miRNA*), a short double-stranded RNA (dsRNA) and a mature miRNA (Bartel, 2004). Finally, mature miRNAs are predominantly incorporated in the RNA-induced silencing complex (RISC) (Bartel, 2004) in which they negatively regulate gene expression by

inhibiting gene translation or degrading coding mRNAs by perfect or near-perfect complement to target mRNAs (Carrington and Ambros, 2003). In animals, target mRNAs usually contain multiple weakly miRNA-complementary sites located at the 3' untranslated regions (UTR); miRNAs imperfectly complement to these sites and possibly hamper ribosome movement along the mRNA, and repress gene expression (Carrington and Ambros, 2003). In plants, most target mRNAs only contain one single miRNA-complementary site, and most corresponding miRNAs perfectly complement these sites and cleave the target mRNAs (Kidner and Martienssen, 2005). However, some miRNAs, such as miRNA 172, regulate gene expression by repressing gene translation, although they can perfectly complement the target mRNAs (Aukerman and Sakai, 2003; Chen, 2004).

Although miRNAs are small, they have versatile functions (Bartel and Bartel, 2003; Hunter and Poethig, 2003; Kidner and Martienssen, 2005; Zhang *et al.*, 2006a). Several experimental and genetic analyses have indicated that microRNAs are essential for plant development. DCL is a key enzyme for processing pri-miRNA to pre-miRNA, then to the miRNA: miRNA* duplex. Loss-of-function *dcl1* mutants of *Arabidopsis thaliana* resulted in decreasing miRNA levels and ectopically increased the expression of miRNA target genes (Kasschau *et al.*, 2003; Park *et al.*, 2002; Reinhart *et al.*, 2002). This caused many developmental deficiencies, such as delaying flower timing, over-proliferation of shoot meristems and embryogenic suspensor cells, and converting normally determinate floral meristems to indeterminate meristems (Carrington and Ambros, 2003). The miRNA miR-JAW is crucial for leaf development. miR-JAW mutants caused uneven leaf shape and leaf curvature in *Arabidopsis thaliana* and in maize by decreasing a subset of TCP (TBI, CYC, PCF) gene expression (Llave *et al.*, 2002; Palatnik *et al.*, 2003). miRNA 164 regulates *CUP-SHAPED COTYLEDONS* (*CUC*) gene expression. Its dysfunction caused aberrant leaf morphology and increasing separation (Laufs *et al.*, 2004). miRNA 172 regulates *APETELA* (*AP1* and *AP2*) gene expression. This miRNA functions in productive translation and causes floral organ identity defects and flowering time alternatives (Aukerman and Sakai, 2003; Chen, 2004). miRNA genes are also involved in hormone signaling (Eckardt, 2005; Guo *et al.*, 2005; Inukai *et al.*, 2005; Mallory *et al.*, 2005) and environmental stress (Jones-Rhoades and Bartel, 2004; Sunkar and Zhu, 2004; Zhang *et al.*, 2005). Several studies have shown that auxin response factor (ARF) genes are targeted by miRNAs (Eckardt, 2005; Guo *et al.*, 2005; Inukai *et al.*, 2005; Mallory *et al.*, 2005), and miRNA 159 is positively regulated by gibberellic acid (GA) (Achard *et al.*, 2004). Cold, heat, pathogens and salinity affect the expression of miRNAs (Sunkar and Zhu, 2004).

Since the first miRNA (*lin-4*) was identified in *Caenorhabditis elegans* in 1993 (Lee *et al.*, 1993; Wightman *et al.*, 1993), hundreds of miRNAs have been discovered in animals, plants and viruses. At present, the sequences of 1650 miRNAs are deposited in the miRNA Registry Database (version Rfam 6.0, released April 2005) (Griffiths-Jones, 2004). Of those, 391 are plant miRNAs, of which 114 are from *Arabidopsis thaliana*, 173 from rice (*Oryza sativa*), 64 from sorghum (*Sorghum bicolor*) and 40 from maize (*Zea mays*). Although some miRNAs were directly isolated and cloned by genetic or biochemical approaches, most plant miRNAs were predicted by computational approaches and then validated by molecular techniques such as Northern analysis or detecting the corresponding clones (Lai, 2003). Although traditional computational approaches have certain advantages and have made great progress in predicting new potential miRNAs, it is difficult to predict miRNAs in species with unsequenced genomes because these approaches are

based on the availability of a genomic sequence. At present, the genomic sequences are available for only for a few plant species. Thus, the majority of investigations on miRNAs are currently limited to a few model species, such as *C. elegans*, human, *Arabidopsis thaliana*, rice, etc. In addition, the traditional computational approach was slightly inefficient and certainly not comprehensive enough to demonstrate the expression of predicted miRNAs.

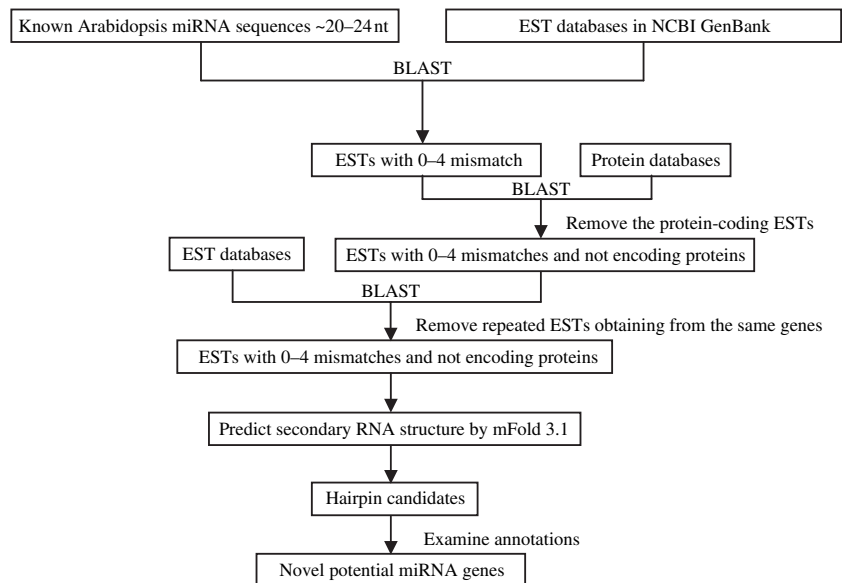
Many miRNAs are evolutionarily conserved from species to species within the same kingdom. miRNA genes in one species may exist as orthologs or homologs in other species (Weber, 2005). This suggests a powerful strategy to identify new miRNA genes. Weber (2005) found 35 human and 45 mouse new potential miRNAs by a homology search. miRNAs are non-coding mRNA, so they may be present in expressed sequence tags (ESTs). Thus, we can search the homologs of known miRNAs in the EST database to discover new miRNAs in other species. Recently, we developed a successful strategy to predict new potential miRNAs in plant species by mining out undiscovered miRNAs from the repository of ESTs currently available (Zhang *et al.*, 2005). ESTs in publicly available databases are not equally and uniformly sampled in all plant species or all plant tissues; this bias possibly influences the results of any data mining effort. However, mining EST databases in a systematic way could provide deeper insight into the distribution and conservation of plant miRNAs, allowing the development of new experimental approaches for understanding miRNA origins, evolution and divergence. Here, we report on the results of a systematic search for potential new plant homologs of *Arabidopsis thaliana* miRNAs deposited in the miRNA Registry Database (version Rfam 6.0, released April 2005) (Griffiths-Jones, 2004) using EST analysis. Using this new strategy, we addressed the following questions. (i) Are miRNA sequences present in EST databases? (ii) Are these miRNAs widely distributed in the plant kingdom or only limited to a subset of taxa closely related to *Arabidopsis thaliana*? (iii) Are some families of miRNAs restricted in their phylogenetic distribution? (iv) Is there evidence for strong or weak evolutionary conservation of miRNA structure and thus function?

Results and discussion

New plant homologs of miRNAs

After searching the EST database and comparing with characteristics of previously known *Arabidopsis thaliana* miRNAs, a total of 670 ESTs were identified that contained the precursor or mature sequences of potential miRNA homologs. Figure 1 summarizes the major steps for identifying EST sequences and their corresponding miRNA sequences. According to the biogenesis and expression criteria suggested by Ambros *et al.* (2003), these ESTs code

Figure 1. Schematic representation of the miRNA gene search procedure used to identify homology of known *Arabidopsis thaliana* miRNA genes.



481 miRNAs, representing 37 miRNA families in 71 plant species (Table 1).

Our newly identified plant miRNA precursors have negative folding free energies (30–160 kcal mol⁻¹ with an average of about -52 kcal mol⁻¹) according to MFOLD (Mathews *et al.*, 1999); this is similar to the computational values of *Arabidopsis thaliana* miRNA precursors (-57 kcal mol⁻¹) and much lower than folding free energies of tRNA (-27.5 kcal mol⁻¹) or rRNA (-33 kcal mol⁻¹) calculated by Bonnet *et al.* (2004b). The foldback precursors of these potential miRNAs are usually about 60–180 nucleotides. The predicted secondary structures (Mathews *et al.*, 1999) indicated that there are at least 16 nucleotides engaged in Watson–Crick or G/U base pairings between the mature miRNA and the opposite arms (miRNA*) in the hairpin structure, and the stem–loop precursor did not contain large internal loops or bulges (Figure 2). Our method identified new potential miRNAs based on searching homologies of *Arabidopsis thaliana* miRNAs in other plant species. This satisfies the biogenesis criterion proposed by Ambros *et al.* (2003). ESTs are partial cDNA sequences of expressed genes (Adams *et al.*, 1991; Matukumalli *et al.*, 2004); this indicates that the predicted miRNA homologs satisfy miRNA expression criteria A and B described by Ambros *et al.* (2003). Thus, at least three criteria were satisfied in miRNAs predicted by this method.

The number of miRNAs obtained by EST analysis depends on three factors: number of previously known miRNAs, conservation of miRNA sequence and structure, and the number of ESTs in the database. We found different numbers of miRNAs in different plant species due to the different number of ESTs in the GenBank EST databases (Table 1). The number of miRNAs was linearly related to the number of ESTs (Figure 3); about 10 000 ESTs contain one

miRNA. In the EST databases, wheat, maize, barley, soybean, rice and *Arabidopsis thaliana* have the greatest numbers of ESTs in the subgroup of Viridiplantae, and all of these plant species have more than 200 000 EST sequences in GenBank's EST databases. In this study, we identified 57, 40, 31, 30, 24 and 21 miRNA homologs, belonging to 20, 20, 14, 15, 10 and 11 miRNA families, for soybean, rice, wheat, maize, barley, and *Arabidopsis thaliana*, respectively. The computational strategy used in recent investigations suggests that miRNAs constitute nearly 1% of predicted protein-coding genes (Lai *et al.*, 2003; Lim *et al.*, 2003a,b). These identified numbers in our study and other studies are very small, indicating that many miRNAs remain to be discovered. Currently, 114 miRNAs have been identified in *Arabidopsis thaliana*, but our EST analysis only detected 21. Of the 418 563 ESTs publicly available, only 0.008% (32 ESTs) contain miRNAs. This rate is much lower than the miRNA constitution rate of predicted protein-coding genes. This discrepancy may be due to two non-mutually exclusive reasons. First, miRNAs or miRNA precursors are usually very short (up to 200 nucleotides), but ESTs usually contain more than 300 nucleotides. Thus, a majority of miRNAs or miRNA precursors cannot be cloned to ESTs. Second, a majority of primary miRNAs are rapidly processed in the nucleus, so miRNA precursors have a lower probability of being cloned to ESTs.

After searching the miRNA Registry Database (version Rfam 6.0, released April 2005) (Griffiths-Jones, 2004), we found that some miRNAs identified by EST analysis were identical to those found by direct cloning or by traditional computational approaches and deposited in the miRNA database. For example, *Arabidopsis thaliana* miRNA 414 was found in three ESTs (BP664132, BP660114 and AU226450). Rice miRNA 413 exists in EST CB683283. This

Table 1 ESTs matched by known *Arabidopsis thaliana* miRNAs and identified as new miRNAs in plants

Plant species	Common name	Group	Family	EST number in dbEST ^a	EST contigs identified with known miRNAs	Identified percentage	Number of miRNAs	Number of miRNA families
<i>Acor americanus</i>	Sweetflag	Monocot	Acoraceae	5825	1	0.017	1	1
<i>Aegilops speltoides</i>		Monocot	Poaceae	4315	1	0.023	4	1
<i>Allium cepa</i>		Monocot	Liliaceae	19 582	14	0.071	7	2
<i>Arabidopsis thaliana</i>		Dicot	Brassicaceae	418 563	32	0.008	21	11
<i>Avena sativa</i>	Oat	Monocot	Poaceae	7624	1	0.013	1	1
<i>Beta vulgaris</i>	Beet	Dicot	Chenopodiaceae	22 706	1	0.004	1	1
<i>Brassica napus</i>	Oilseed rape	Dicot	Brassicaceae	63 558	3	0.005	3	3
<i>Capsicum annuum</i>		Dicot	Solanaceae	30 229	3	0.010	3	2
<i>Citrus sinensis</i>		Dicot	Rutaceae	87 620	5	0.006	5	3
<i>Citrus unshiu</i>		Dicot	Rutaceae	2561	1	0.039	1	1
<i>Cycas rumphii</i>		Gymnosperm	Cycadaceae	5952	1	0.017	1	1
<i>Eschscholzia californica</i>		Dicot	Papaveraceae	9083	1	0.011	1	1
<i>Glycine clandestina</i>		Dicot	Fabaceae	933	2	0.214	2	2
<i>Glycine max</i>	Soybean	Dicot	Fabaceae	355 970	70	0.020	57	20
<i>Glycine soja</i>		Dicot	Fabaceae	16 508	3	0.018	1	1
<i>Gossypium arboreum</i>		Dicot	Malvaceae	39 216	5	0.013	5	5
<i>Gossypium hirsutum</i>		Dicot	Malvaceae	24 050	1	0.004	1	1
<i>Gossypium raimondii</i>	Upland cotton	Dicot	Malvaceae	63 577	18	0.028	12	7
<i>Hedyotis centranthoides</i>		Dicot	Rubiaceae	5416	3	0.055	3	3
<i>Hedyotis terminalis</i>		Dicot	Rubiaceae	4875	1	0.021	1	1
<i>Helianthus annuus</i>		Dicot	Asteraceae	66 098	4	0.006	4	4
<i>Hordeum vulgare</i>	Barley	Monocot	Poaceae	394 937	39	0.010	24	10
<i>Ipomoea batatas</i>		Dicot	Convolvulaceae	4200	1	0.024	1	1
<i>Ipomoea nil</i>		Dicot	Convolvulaceae	25 946	10	0.039	5	4
<i>Lactuca sativa</i>		Dicot	Asteraceae	68 774	8	0.012	6	6
<i>Liriodendron tulipifera</i>		Dicot	Magnoliaceae	6255	2	0.032	2	2
<i>Lotus corniculatus</i> var. <i>japonicus</i>		Dicot	Fabaceae	111 623	7	0.006	6	5
<i>Lupinus luteus</i>		Dicot	Fabaceae	364	1	0.275	1	1
<i>Lycopersicon esculentum</i>	Tomato	Dicot	Solanaceae	197 792	13	0.007	9	5
<i>Malus x domestica</i>	Apple tree	Dicot	Rosaceae	183 937	5	0.003	4	3
<i>Medicago truncatula</i>	Barrel medic	Dicot	Fabaceae	216 703	19	0.009	13	8
<i>Mesembryanthemum crystallinum</i>	Common ice plant	Dicot	Aizoaceae	25 803	7	0.027	2	1
<i>Nicotiana benthamiana</i>		Dicot	Solanaceae	26 986	6	0.022	5	3
<i>Nicotiana tabacum</i>	Tobacco	Dicot	Solanaceae	27 054	1	0.004	1	1
<i>Nuphar advena</i>		Dicot	Nymphaeaceae	8435	3	0.036	3	1
<i>Oryza sativa</i>	Rice	Monocot	Poaceae	406 624	57	0.014	40	20
<i>Pennisetum ciliare</i>	Buffelgrass	Monocot	Poaceae	987	1	0.101	1	1
<i>Pennisetum glaucum</i>		Monocot	Poaceae	2532	1	0.039	1	1
<i>Persea americana</i>		Dicot	Lauraceae	7883	1	0.013	1	1
<i>Phaseolus coccineus</i>		Dicot	Fabaceae	20 120	2	0.010	2	2
<i>Physcomitrella patens</i>		Moss	Funariaceae	82 420	2	0.002	2	2
<i>Picea engelmannii</i> × <i>Picea sitchensis</i>		Gymnosperm	Pinaceae	12 127	3	0.025	3	3
<i>Picea glauca</i>		Gymnosperm	Pinaceae	54 819	11	0.020	11	8

Table 1 Continued

Plant species	Common name	Group	Family	EST number in dbEST ^a	EST contigs identified with known miRNAs	Identified percentage	Number of miRNAs	Number of miRNA families
<i>Picea sitchensis</i>	Sitka spruce	Gymnosperm	Pinaceae	12 065	5	0.041	4	4
<i>Pinus pinaster</i>		Gymnosperm	Pinaceae	18 254	1	0.005	1	1
<i>Pinus taeda</i>	Loblolly pine	Gymnosperm	Pinaceae	231 685	15	0.006	7	5
<i>Populus alba</i> × <i>Populus tremula</i>		Dicot	Salicaceae	7595	3	0.039	3	2
<i>Populus balsamifera</i> subsp. <i>trichocarpa</i>		Dicot	Salicaceae	58 146	8	0.014	7	6
<i>Populus balsamifera</i> × <i>Populus deltoides</i>		Dicot	Salicaceae	33 134	5	0.015	5	3
<i>Populus euphratica</i>		Dicot	Salicaceae	13 903	5	0.036	5	2
<i>Populus tremula</i>		Dicot	Salicaceae	37 313	7	0.019	12	4
<i>Populus tremula</i> × <i>Populus tremuloides</i>		Dicot	Salicaceae	76 160	19	0.025	8	8
<i>Prunus armeniaca</i>		Dicot	Rosaceae	15 081	2	0.013	2	2
<i>Prunus persica</i>		Dicot	Rosaceae	21 873	1	0.005	1	1
<i>Robinia pseudoacacia</i>		Dicot	Fabaceae	2933	3	0.102	1	1
<i>Saccharum officinarum</i>		Monocot	Poaceae	246 301	45	0.018	32	14
<i>Saccharum sp</i>		Monocot	Poaceae	9636	1	0.010	2	2
<i>Schedonorus arundinaceus</i>		Monocot	Poaceae	2462	2	0.081	1	1
<i>Secale cereale</i>		Monocot	Poaceae	9196	2	0.022	1	1
<i>Sesamum indicum</i>		Dicot	Pedaliaceae	3328	1	0.030	2	2
<i>Solanum tuberosum</i>	Potato	Dicot	Solanaceae	214 382	3	0.001	17	12
<i>Sorghum bicolor</i>	Sorghum	Monocot	Poaceae	208 197	21	0.010	17	12
<i>Sorghum propinquum</i>		Monocot	Poaceae	21 780	35	0.161	2	2
<i>Theobroma cacao</i>		Dicot	Sterculiaceae	6572	2	0.030	1	1
<i>Triticum aestivum</i>	Wheat	Monocot	Poaceae	589 476	49	0.008	31	14
<i>Triticum turgidum</i>		Monocot	Poaceae	8714	1	0.011	1	1
<i>Vitis vinifera</i>		Dicot	Vitaceae	147 300	8	0.005	7	6
<i>Zea mays</i>	Maize	Monocot	Poaceae	452 984	52	0.011	30	15
<i>Zinnia elegans</i>		Dicot	Asteraceae	17 529	3	0.017	3	3
Total				5 606 581	670	0.012	481	37

^adbEST release 060305: total 27 296 775 EST entries, summary by organism at 3 June 2005.

(a) miRNA 398

Citrus unshiu
 A A G G .-UAAUUUGAUUU UU U
 GAGG GUGA CCUGAGAACA AGGGU CGUUGGU GCA GC G
 UUCC CACU GGACUCUUGU UCCAC GCAACCGG CGU CG C
 ^ C - G - (34 nt side loop) C- U

Glycine max
 A A CU UG A GC UC
 GAGG GUGA UCUGAGAACAACAGG GGUU C CU UUAUACA \\
 UUCC CACU GGACUCUUGUUGU UUA G GA AUAUGGU U
 ^ C - AU GU - -- UA

Gossypium raimondii
 CC A UC CCCUU
 GAGGGGUG ACCU AGAACACAGG AU \\
 UUCCCCAC UGGA UCUUGUGUUC UA A
 ^ -- C GA AACCA

Lactuca sativa
 U A C UUA- CAA
 CAGG GCGAC UGG AACACAUG AAUGU C
 GUCC CCGUG ACU UUGUGUAC UUACAC A
 ^ C G C CGCC UAA

Lotus corniculatus var. japonicus (Lotus japonicus)
 A G C UUGAAUU- - C C
 AAGG GUGA CCUGAGAACAACAG UGAAUUGU GCC AUAUCA ADA U
 UUCC CACU GGACUCUUGUUGU AUUUAUUG CGG UAUGU UAU G
 ^ C - U UUGUAUUU A C A

Medicago truncatula
 U A AGCUAU-- U
 CAGGG CGAC UGAGAGCACAUCA CAUGG U
 GUCCC CGUC ACUCUUGUUAU GUAUC G
 ^ C G CAACCUAAU U

Nicotiana benthamiana
 A U U UCAUUUUUUUUU U UC
 CAGGGGC ACCUGAGA CACAUA UG AG UGUUGAG \\
 GUCCCCG UGGACUCU GUGUAU AC UC AUAACUU U
 ^ C U C ----- U GG

Oryza sativa
 G U- G GC
 CAGGGGCGA CUGGGAACACACCG GAU AG \\
 GUCCCCGCU GACUCUUGUUGU CUG UC G
 ^ G U U G UG

Populus euphratica
 A G C---- U
 CAGG GCGACCU GAAUACAUGU GCUC ACCC C
 GUCC CCGUGGAC CUUGUGUACA CCGG UGGG U
 ^ C U A- UUCUCU U

(b) miRNA 408

Arabidopsis thaliana
 A CAA A AUU UUU U UAAAA
 G CAGGGAA GCAG GCAUGG GAG AC AAAACA \\
 C GUCCCUU CGUC CGUACC CUC UG UUUUGU C
 G^ CUC A CAU --- - CUCAG

Arabidopsis thaliana
 A CAA A .-AUUGA A
 G CAGGGAA GCAG GCAUGG GUUU C
 C GUCCCUU CGUC CGUACC CAAA U
 G^ CUC A CAU (26 nt side loop) A

Glycine max
 A C A GA U CAA G- AAGAA UG
 GC GGGGAA AGGCAG GCAUG UGGAGCUA CAACA UAUU UC AC \\
 CG UCCCUU UCCGUC CGUAC GUCUUGGU GUUGU AUAA AG UG A
 ^ G C A UC - --- AG GAGAG AG

Saccharum officinarum
 A U A U CAU .-AAAUUUCCAAUUUCUGUC CGC- C UG .-A CUUG-- ACCG
 G CAGGGAA GAGGCAG GCA GGA GAGGC CAACA CUC UAGGCCG UAC C UUUCUGUUUG CUCACAAA \\
 C GUCCCUU CUCCGUC CGU CCU CUUCG GUUGU GAG GUCCGUG GUG G AAAGACGGAC GAGUGUUU A
 G^ U U A C C UU- \ (25 nt side loop)- UCCA A GU (27 nt side loop) UAUUGA AGGG

Triticum aestivum
 - - - U A G - AA-- AAAA A AC GAUU
 G AGGGGGAGG AG CAGGGA GGAGCAG AGCA G GAU GAGGC GCAAC UUU CC CU A
 C UCCCUUCC UC GUCCU CCUCGUGU UUGU C UUG CUUCG CGUUG AGA GG GA U
 GG^ G AC C G - G U AGAC AG-- - GA AGAG

Zea mays
 A U U U CAU AA .-GG CUA UGUGA GAAA
 G CAGGGAA GAGGCAG GCA GGA GGGC CAACA GU AGGGA GCU GGCA \\
 C GUCCCUU CUCCGUC CGU CCU CUUCG GUUGU CG UCUCU CGG CCGU G
 G^ U A C C UU- GA (15 nt side loop) --- UGA-- GGAA

Oryza sativa
 A U A U UAU .-AUGUAG UC G
 G CAGGGA GAGGCAG GCA GGA GGGC CAACAG AUUAU CUU C
 C GUCCCUU CUCCGUC CGU CCU CUUC GUUGUU UAGUA GAA A
 G^ U A C C UU- (36 nt side loop) GA C

Populus balsamifera subsp. trichocarpa x Populus deltoides
 A C A A CU GAA
 G CAGGGAA AGGCAG GCAUGG UGGAGCUA AAAC G
 C GUCCCUU UCCGUC CGUACC AUUCUGGU UUGU U
 G^ C A C U- ACA

Sorghum bicolor
 C U A U .-AUG A .-ACAAAAUU AAU GCUUG
 G CAGGGA GA GCAG GCA GGG GGGCCAUA CA UCC UCCCGUUU \\
 C GUCCCUU CU CGUC CGU CCC UCCGGUAGU GU AGG GAGGUAAA C
 G^ U C A - (39 nt side loop) G (48 nt side loop) --- ACACC

Figure 2. Representative miRNAs found in this study in different plant species. The miRNA sequence is shown by letters in grey. (a) miRNA 398 in nine different plant species. (b) miRNA 408 in nine different plant species.

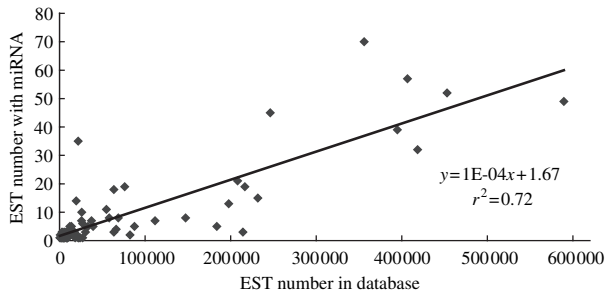


Figure 3. The linear relationship between the numbers of miRNA found in the EST database and the numbers of ESTs for certain plant species.

indicates that EST analysis is a novel alternative approach to identify miRNAs. EST analysis makes it possible to rapidly study miRNAs and their functions in species for which the genome sequences are not well known, which would be impossible using traditional computational approaches. However, EST analysis cannot identify rapidly evolving non-conserved miRNAs. Non-conserved miRNAs are usually species-specific and are likely to be very abundant in organisms including plants. However, EST analysis provides a cost-effective and rapid route towards the discovery and isolation of conserved miRNAs. As more ESTs become available in public databases and more miRNAs are identified in genome-sequencing model species, the likelihood of successfully mining EST databases will increase.

Lohmann and Weigel, 2002). Recent studies indicate that these class A genes are targets of miRNA 172 in *Arabidopsis thaliana*, and over-expression of miRNA 172 inhibited the translation of the *ap2* and *toe1* genes and resulted in early flowering and disruption of the specification of floral organ identity similar to loss-of-function *ap2* gene mutants (Aukerman and Sakai, 2003; Chen, 2004). In this study, we found miRNA 172 in 19 plant species. If one adds these to the other five plant species in which miRNA 172 has been found, the total number of plant species becomes 24, representing 12 plant families. After searching the EST and protein databases, we found at least 17 plant species with potential targets of miRNA 172. The majority of these targets belong to *ap2* gene families. They all perfectly or near perfectly complement miRNA 172. Of these genes, AP2-like gene *glossy15 (gl15)* has been confirmed by one recent study in maize (Lauter *et al.*, 2005). The presence of targets in the EST and protein databases provides additional evidence for the real existence of these EST-predicted miRNAs in the respective plant species.

Interestingly, some ESTs start or end the several nucleotides up or down from the miRNA–mRNA complementary sites. This may simply be by chance; however, it may be some evidence of miRNA cleavage of target mRNAs.

The diversity of miRNAs in plants

While animal miRNA precursors typically have 70–80 nucleotides, plant miRNA precursors are more diverse in structure and size (Figure 5). They varied in size from 60 to 509 nucleotides, with an average of 144.6 ± 56.9 ($n = 513$); most (73.5%) of the detected miRNAs had 81–160 nucleotides. Only 1.6% of plant miRNAs were less than 81 nucleotides in length, a stark contrast to animal miRNAs. This difference in length between plant and animal miRNAs makes it difficult to predict new plant miRNAs using the same computational approaches. Different sizes of miRNA precursors usually produce a slightly different secondary stem–loop hairpin structure, although this structure is conserved within the same miRNA family (Figure 2). The longer precursors may host other important functional elements and offer unique functions for regulating miRNA biogenesis or gene expression.

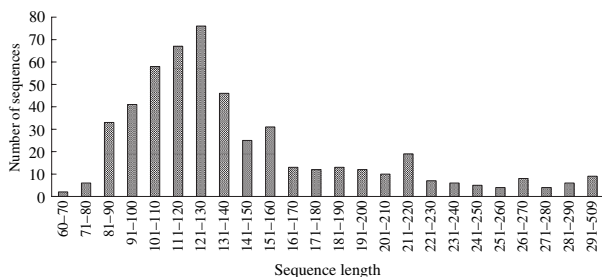


Figure 5. Size distribution of previously known plant miRNAs.

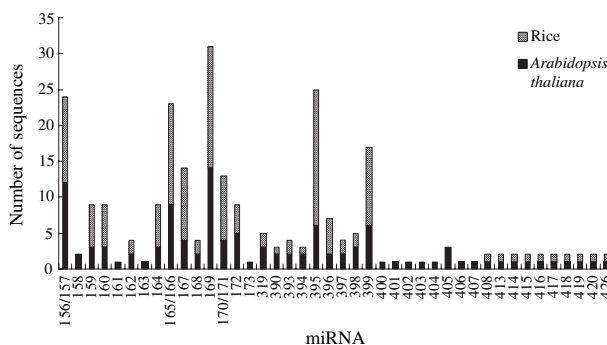


Figure 6. miRNA family size in *Arabidopsis thaliana* and rice.

To date, a total of 117 miRNAs have been identified from *Arabidopsis thaliana*. They belong to 43 miRNA families based on their sequence similarity. Different miRNA families have different sizes (Figure 6). Some miRNA families have more than 10 members, such as miRNA 156/157, miRNA 169 and miRNA 395. However, only one or two members were identified for the majority of the plant miRNA families. The size of miRNA families may be indicative of their function.

The conservation and divergence of miRNAs in plants

In this study, we not only identified 481 miRNAs in 71 different plant species, but also found solid evidence that miRNAs are highly conserved in the plant kingdom, irrespective of the time of evolutionary divergence. Eighteen families of miRNAs have orthologs or homologs in more than 10 different plant species, spanning the breadth of green plant phylogeny (Table 3). Of these miRNAs, miRNA 156/157, miRNA 172 and miRNA 170/171 have orthologs in 45, 24 and 22 different plant species, which belong to 21, 12 and 12 plant families, respectively (Table 3). The miRNAs found in more than 10 different plant families should be considered as highly conserved miRNAs. In addition to these three miRNA families, miRNA 165/166, miRNA 159/319, miRNA 396, miRNA 168, miRNA 160 and miRNA 390 are also found in at least 10 plant families; these six miRNA families are also considered as highly conserved miRNAs. Nine miRNA families (miRNAs 394, 164, 169, 167, 162, 398, 414, 393, 397 and 163) were found in 5–9 different plant families. We classified these miRNA families as moderately conserved miRNAs. This suggests that these miRNAs play important and conserved functions in plant development, such as flower and leaf development. We also found that some miRNAs are low conserved or non-conserved in plants (Table 3). These low or non-conserved miRNAs may play roles in more species-specific characteristics in plant development, such as cotton fiber differentiation, elongation and development. In a previous study, miRNAs 163 and 158 were identified as non-conserved miRNAs (Jones-Rhoades and Bartel, 2004). After EST analysis, five and four plant species

Table 3 Sensitivity and conservation of miRNAs in plants

miRNA family	Total number of plants found miRNAs in their ESTs		Number of plant species (families) recorded miRNAs in the miRNA Registry and other studies	Total number of plants in which miRNAs found	
	Number of plant species	Number of plant families		Number of plant species	Number of plant families
Highly conserved miRNAs					
156/157 ^a	39	17	15 (10)	45	21
172 ^b	19	9	10 (8)	24	12
170/171 ^{a,c}	16	9	11 (8)	22	12
165/166 ^a	11	5	13 (9)	18	11
159/319 ^{a,b}	17	7	14 (9)	21	10
396 ^d	15	9	11 (7)	21	10
168 ^a	15	7	12 (8)	20	10
160 ^a	8	4	13 (9)	18	10
390 ^e	5	5	8 (7)	12	10
Moderately conserved miRNAs					
394 ^d	16	7	8 (5)	19	9
164 ^a	5	5	7 (6)	12	9
169 ^a	7	3	12 (8)	15	8
167 ^{a,b,c}	9	4	11 (7)	14	8
162 ^{a,c}	7	5	6 (5)	11	8
398 ^{d,e}	9	7	4 (4)	11	8
414 ^f	11	7	2 (2)	11	7
393 ^{d,e}	7	4	6 (4)	12	6
397 ^{d,e}	6	5	5 (4)	11	6
163 ^a	7	5	2 (2)	7	5
Lowly conserved miRNAs					
395 ^{d,e}	4	2	6 (4)	9	4
408 ^e	8	4	5 (3)	9	4
399 ^d	4	3	6 (4)	8	4
158 ^a	4	3	2 (2)	6	4
403 ^e	4	3	2 (2)	5	4
161 ^a	2	2	2 (2)	4	4
406 ^e	7	2	1 (1)	8	3
173 ^b	1	1	3 (3)	3	3
419 ^f	2	1	2 (2)	4	2
415 ^f	1	1	2 (2)	3	2
413 ^f	1	1	2 (2)	2	2
416 ^f	0	0	2 (2)	2	2
417 ^f	0	0	2 (2)	2	2
418 ^f	0	0	2 (2)	2	2
420 ^f	0	0	2 (2)	2	2
426 ^f	0	0	2 (2)	2	2
Non-conserved miRNAs					
400 ^e	0	0	1 (1)	1	1
401 ^e	0	0	1 (1)	1	1
402 ^e	0	0	1 (1)	1	1
404 ^e	0	0	1 (1)	1	1
405 ^e	0	0	1 (1)	1	1
407 ^e	0	0	1 (1)	1	1

All previously known miRNAs were tallied. The number of plant species was found by searching the EST database and predicting the secondary structures of ESTs, and from the miRNA Registry Database (Griffiths-Jones, 2004) and other studies. Citations for previously identified miRNAs: ^aReinhart *et al.* (2002); ^bPark *et al.* (2002); ^cLlave *et al.* (2002); ^dJones-Rhoades and Bartel (2004); ^eSunkar and Zhu (2004); ^fWang *et al.* (2004).

have been identified as having homologs of these two miRNAs, respectively. They could actually be classified as moderately or low conserved miRNA families.

There are two methods for identification of miRNAs: the cloning approach and traditional computational approach.

Cloning is not highly efficient at finding all miRNAs due to the low expression levels of miRNAs. Traditional computational approaches usually require genome sequences, which are unavailable for a majority of plant species. This makes it difficult to draw the correct conclusion, such as in a recent

study where Axtell and Bartel (2005) did not find miRNA 162 in all tested plant species including *Arabidopsis thaliana*. However, we found seven plant species with miRNA 162 using EST analysis. Thus, mining EST databases in a systematic way could overcome some of the shortcomings of currently used cloning and computational approaches for identifying new miRNAs, and provide deeper insight into the distribution and conservation of plant miRNAs. ESTs are created unequally from different tissues and different plant species. This sampling problem may affect the ability to evaluate miRNA conservation in plants in closely related plant groups, such as in the same plant family. This should be kept in mind when comparing the distribution and conservation of miRNAs in a small group of plant species.

Of 292 previously known *Arabidopsis thaliana* and rice mature miRNAs, 220 (83.6%) have U at the first position from the 5' end (Figure 7). Less than 25% of other positions in this orientation contained a uracil. This may be a unique characteristic of plant miRNAs, and may play some important role in mature miRNA biogenesis or RISC formation.

Plant miRNAs are highly conserved among distantly related plant species (Figures 8 and 9), both in terms of primary and mature miRNAs, but especially for mature sequences and their complementary miRNA* sequence. Even for comparisons between mosses and core eudicots (*Physcomitrella patens*, *Arabidopsis thaliana* and rice), the mature RNA sequences are almost identical (Figure 9). Some miRNA families do appear to be restricted to particular plant groups; for example, phylogenetic analysis of the miRNA 156 family indicates that sequences present in monocots (such as rice, maize, sorghum) are divergent from sequences present in dicot species (such as *Arabidopsis thaliana*) (Figure 8a). This result solidly confirms that plant miRNAs are evolutionarily conserved.

Several miRNA families have multiple members within the same plant species. For instance, miRNA 395 has 18 members in rice. Although they are conserved as mature miRNA and miRNA* sequences, the other parts of miRNA precursor differ widely, from being closely related to being rather distantly related (Figure 10). These results suggest that the different members of the same miRNA family may evolve at different rates within the same plant species.

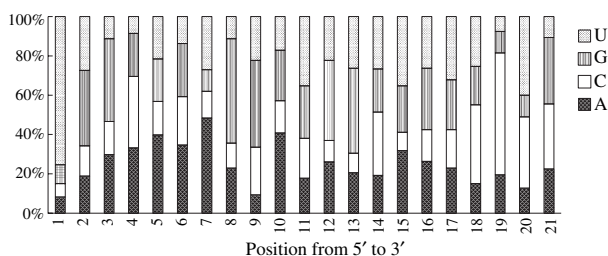


Figure 7. Position-specific nucleotide preferences in plant miRNAs.

The origin of plant miRNAs

In previous studies, miRNAs were considered to be conserved in plants based on limited data and evidence. Most miRNAs of dicots (based on *Arabidopsis thaliana*) have homologs in monocots (primarily based on rice) (Adai *et al.*, 2005; Bonnet *et al.*, 2004a; Jones-Rhoades and Bartel, 2004; Llave *et al.*, 2002; Reinhart *et al.*, 2002; Sunkar and Zhu, 2004; Wang *et al.*, 2004). It is thought that some miRNAs diverged more than 125 million years ago. Recently, Axtell and Bartel (2005) employed microarray technology to detect miRNAs in different plant species. They observed that miRNA159/319 was expressed in 10 different plant species including a moss species; miRNA 156/157 and miRNA 165/166 were expressed in nine plant species. To study miRNA conservation among different species, especially among distantly related species, more species need to be surveyed. However, no study to date has shown that the same miRNA exists in more than 10 plant species.

Identical miRNA sequences exist in closely and distantly related plant species, with little apparent effect of phylogenetic distance. Many of the miRNAs discovered were found in a wide range of plant groups, from mosses to angiosperms. Some miRNA families exist broadly within the angiosperms, including eudicots and monocots, dating back to at least the early Cretaceous (Figure 11). Several miRNA families also pre-date the divergence of gymnosperms and angiosperms (305 million years) and the divergence between vascular plants and mosses (490 million years). These results indicate that miRNA sequences are highly conserved across great phylogenetic distances and that similar selection pressures have been active in the regulation of gene expression in plant cells since the earliest stages of their evolution.

Not only are miRNA genes conserved across all plant lineages, but we also observed that their targets are conserved in different plant families. Although there are many nucleotide changes among the targets of different plant species, the sequences of the complementary sites are highly conserved. This result is similar to observations by Floyd and Bowman (2004). In their study, they observed that the class III homeodomain-leucine zipper (HD-Zip) genes, one of the targets of miRNA 166, have conserved miRNA 166 target regions, although other regions have much lower nucleotide conservation. Based on this finding, they hypothesized that miRNA-mediated mechanisms have existed in plant phylogeny for more than 400 million years. In a broader sense, our results indicate that gene regulation by miRNA is an ancient evolutionary mechanism to control the variety of gene expression. It is one part of the universal gene regulation mechanism. We estimate that miRNA-mediated gene regulation existed more than 425 million years ago in the plant kingdom (Figure 11). The age of plant miRNA is comparable to the age of miRNA regulation in

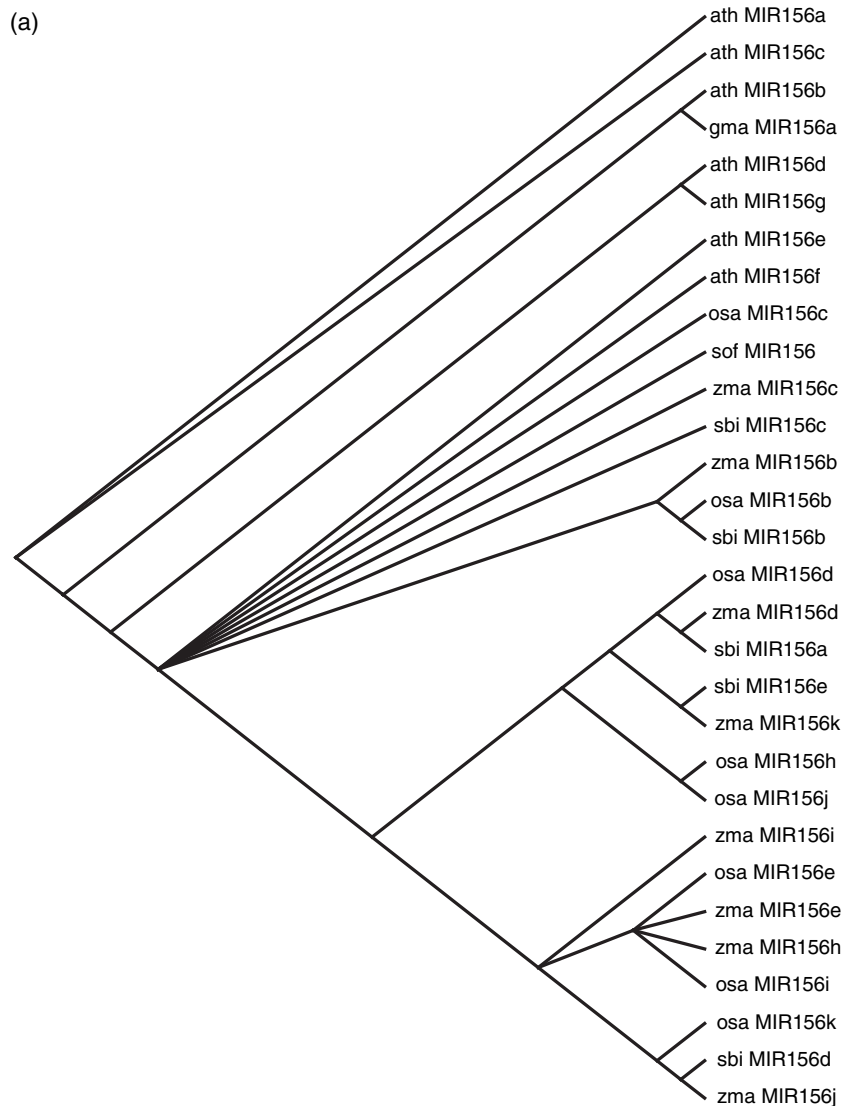


Figure 8. Polygenetic tree (a) and multiple sequence alignment (b) of miRNA 156 from seven plant species.

metazoans (Pasquinelli *et al.*, 2000), and to the age of miRNA targets in plants (Floyd and Bowman, 2004).

Experimental procedures

Identifying potential plant miRNAs using EST analysis and miRNA sequence analysis

The sequences of previously known *Arabidopsis thaliana* and rice mature and precursor miRNAs were obtained from the miRNA Registry Database (version Rfam 6.0, released April 2005) (Griffiths-Jones, 2004). Figure 1 summarizes the major steps for identifying EST sequences and their corresponding miRNA sequences. The mature sequences of 114 previously known *Arabidopsis thaliana* miRNA were subjected to a BLAST search in the sub-group of Viridiplantae of the publicly available EST databases using Blastn 2.2.9 (1 May 2004) (Altschul *et al.*, 1997). Adjusted blast parameter settings were as follows: expect values were set

at 1000; low complexity was chosen as the sequence filter; the number of descriptions and alignments was raised to 1000. The default word-match size between the query and database sequences was 7. If the matched sequences were less than the previously known *Arabidopsis thaliana* mature miRNA sequences, the non-aligned parts were manually inspected and compared to determine the number of matching nucleotides. All BLAST results were saved. EST sequences which closely matched the previously known *Arabidopsis thaliana* miRNAs were manually chosen. Close matching was defined as n/n , $n-1/n$, $n-2/n$ and $n-3/n$ nucleotide matches, where n equals the previously known *Arabidopsis thaliana* miRNA length.

The secondary structures of the selected EST sequences were carefully predicted and generated using the Zuker folding algorithm with MFOLD 3.1 (Mathews *et al.*, 1999; Zuker, 2003) that is publicly available at <http://www.bioinfo.rpi.edu/applications/mfold/old/rna/>. The following parameters were used in predicting the secondary structures: folding temperature was fixed at 37°C; ionic conditions were set at 1 M NaCl and with no divalent ions; and the grid lines in

Figure 9. Comparison of mature miRNAs from different plant species. (a) miRNA 156; (b) miRNA 172; (c) miRNA 319; (d) miRNA 396.

(a) miRNA 156		(b) miRNA 172	
	-----+-----		-----
AtMiR156a	TGACAGAGAGAGTGGGCAC	GrMiR172b	AAAATCTTGATGATGCTGCAT
BnMiR156	TGACAGAGAGAGTGGGCAC	CsMiR172a	AGAATCTTGATGATGCTGCAT
GmMiR156a	TGACAGAGAGAGTGGGCAC	GmMiR172n	AGAATCTTGATGATGCTGCAT
HvMiR156a	TGACAGAGAGAGTGGGCAC	GrMiR172a	AGAATCTTGATGATGCTGCAT
OsMiR156a	TGACAGAGAGAGTGGGCAC	LeMiR172	AGAATCTTGATGATGCTGCAT
SbMiR156a	TGACAGAGAGAGTGGGCAC	StMiR172	AGAATCTTGATGATGCTGCAT
ScMiR156a	TGACAGAGAGAGTGGGCAC	GmMiR172n	AGAATCTTGATGATGCTGCAT
StMiR156a	TGACAGAGAGAGTGGGCAC	CsMiR172b	AGAATCTTGATGATGCTGCAT
ZnMiR156a	TGACAGAGAGAGTGGGCAC	PgMiR172	AGAATCTTGATGATGCTGCAT
AtMiR156b	TGACAGAGAGAGAGGCAC	PpeMiR172	AGAATCTTGATGATGCTGCAT
CaMiR156	TGACAGAGAGAGAGAGGCAC	PsMiR172	AGAATCTTGATGATGCTGCAT
CrMiR156	TGACAGAGAGAGAGAGGCAC	ZnMiR172n	AGAATCTTGATGATGCTGCAT
CsMiR156	TGACAGAGAGAGAGAGGCAC	AtMiR172n	AGAATCTTGATGATGCTGCAT
EcMiR156	TGACAGAGAGAGAGAGGCAC	GhMiR172	AGAATCTTGATGATGCTGCAT
GmMiR156b	TGACAGAGAGAGAGAGGCAC	HvMiR172a	AGAATCTTGATGATGCTGCAT
HaMiR156	TGACAGAGAGAGAGAGGCAC	HvMiR172b	AGAATCTTGATGATGCTGCAT
HvMiR156b	TGACAGAGAGAGAGAGGCAC	HvMiR172c	AGAATCTTGATGATGCTGCAT
LeMiR156	TGACAGAGAGAGAGAGGCAC	HvMiR172d	AGAATCTTGATGATGCTGCAT
LjMiR156	TGACAGAGAGAGAGAGGCAC	MiZnR172n	AGAATCTTGATGATGCTGCAT
LsMiR156	TGACAGAGAGAGAGAGGCAC	OsMiR172n	AGAATCTTGATGATGCTGCAT
MdMiR156	TGACAGAGAGAGAGAGGCAC	VvMiR172	AGAATCTTGATGATGCTGCAT
NbMiR156	TGACAGAGAGAGAGAGGCAC	TcMiR172	AGAATCTTGATGATGCTGCAT
OsMiR156b	TGACAGAGAGAGAGAGGCAC	TaMiR172b	AGAATCTTGATGATGCTGCAT
PaMiR156	TGACAGAGAGAGAGAGGCAC	TaMiR172a	AGAATCTTGATGATGCTGCAT
PgMiR156	TGACAGAGAGAGAGAGGCAC	SoMiR172b	AGAATCTTGATGATGCTGCAT
PpMiR156	TGACAGAGAGAGAGAGGCAC	SoMiR172a	AGAATCTTGATGATGCTGCAT
PptMiR156	TGACAGAGAGAGAGAGGCAC	SbMiR172	AGAATCTTGATGATGCTGCAT
PtMiR156	TGACAGAGAGAGAGAGGCAC	OsMiR172p	AGAATCTTGATGATGCTGCAT
VvMiR156	TGACAGAGAGAGAGAGGCAC	OsMiR172o	AGAATCTTGATGATGCTGCAT
ZnMiR156b	TGACAGAGAGAGAGAGGCAC	OsMiR172n	AGAATCTTGATGATGCTGCAT
SbMiR156b	TGACAGAGAGAGAGAGGCAC	AsMiR172	AGGATCTTGATGATGCTGCAT
ScMiR156b	TGACAGAGAGAGAGAGGCAC	TaMiR172c	AGGATCTTGATGATGCTGCAT
SiMiR156	TGACAGAGAGAGAGAGGCAC	Consensus	AGAATCTTGATGATGCTGCA.
SpMiR156	TGACAGAGAGAGAGAGGCAC		
StMiR156b	TGACAGAGAGAGAGAGGCAC		
TaMiR156	TGACAGAGAGAGAGAGGCAC		
Consensus	TGACAGAGAGAGAGAGGCAC		
(c) miRNA 319		(d) miRNA 396	
ppa-miR319	TTGGACTGAGGGGAGCTCC	gcl-396	TTCCACAGCTTTCTTGAAGCTG
ath-miR319c	TTGGACTGAGGGGAGCTCCT	gna-396a	TTCCACAGCTTTCTTGAAGCTG
vvi-miR319	TTGGACTGAGGGGAGCTCCT	ncr-396a	TTCCACAGCTTTCTTGAAGCTG
ath-miR319a	TTGGACTGAGGGGAGCTCCC	osa-396	TTCCACAGCTTTCTTGAAGCTG
ath-miR319b	TTGGACTGAGGGGAGCTCCC	ppt-396	TTCCACAGCTTTCTTGAAGCTG
gna-miR319	TTGGACTGAGGGGAGCTCCC	ppe-396	TTCCACAGCTTTCTTGAAGCTG
ltu-miR319	TTGGACTGAGGGGAGCTCCC	sof-396	TTCCACAGCTTTCTTGAAGCTG
ptr-miR319	TTGGACTGAGGGGAGCTCCC	zna-396	TTCCACAGCTTTCTTGAAGCTG
tae-miR319a	TTGGACTGAGGGGAGCTCCC	ncr-396b	TTCCACAGCTTTCTTGAAGCTG
osa-miR319a	TTGGACTGAGGGGAGCTCCC	bvu-396	TTCCACAGCTTTCTTGAAGCTG
osa-miR319b	TTGGACTGAGGGGAGCTCCC	ppd-396	TTCCACAGCTTTCTTGAAGCTG
sbi-miR319	TTGGACTGAGGGGAGCTCCC	stu-396	TTCCACAGCTTTCTTGAAGCTG
sof-miR319a	TTGGACTGAGGGGAGCTCCC	ptr-396	TTCCACAGCTTTCTTGAAGCTG
zna-miR319	TTGGACTGAGGGGAGCTCCC	bna-396	TTCCACAGCTTTCTTGAAGCTG
Consensus	TTGGACTGAGGGGAGCTCC.	gna-396b	TTCCACAGCTTTCTTGAAGCTG
		ncr-396c	TTCCACAGCTTTCTTGAAGCTG
		pps-396	TTCCACAGCTTTCTTGAAGCTG
		pgl-396	TTCCACAGCTTTCTTGAAGCTG
		Consensus	TTCCACAGCTTTCTTGAAGCTG.

(ii) a perfect or near-perfect hairpin secondary structure predicted by MFOLD; (iii) a maximum size of 3 nt for a bulge in the miRNA sequence; and (iv) an MFEI greater than 0.85.

Predicted miRNAs and their related information were recorded. Closely related EST sequences were blasted against each other and analyzed. If the ESTs had a high similarity (>98%), it indicated that these ESTs were created from the same mRNA, and these were then considered as one miRNA. Here, we used an identity threshold paired with a minimum overlap length instead of e-values to merge redundant miRNAs, because the e-value is heavily dependent on the sequence length, which is different among different miRNAs.

Multiple sequence alignments of selected miRNA families were performed with the web-based computer software Multalin ([\[prodes.toulouse.inra.fr/multalin/multalin.html\]\(http://prodes.toulouse.inra.fr/multalin/multalin.html\)\) \(Corpet, 1988\). The phylogenetic relationships among these DNA sequences were then reconstructed using the default parsimony analysis in PAUP*, using the heuristic search strategies, random order addition, and 50 repetitions \(Swofford, 2000\). Equally parsimonious trees were combined into a strict consensus tree, collapsing all ambiguous relationships.](http://</p>
</div>
<div data-bbox=)

Potential targets of the EST-predicted miRNAs

To examine whether the potential targets for the EST-predicted miRNAs were present in the given plant species, we also searched the EST and protein-coding gene databases for the complement to

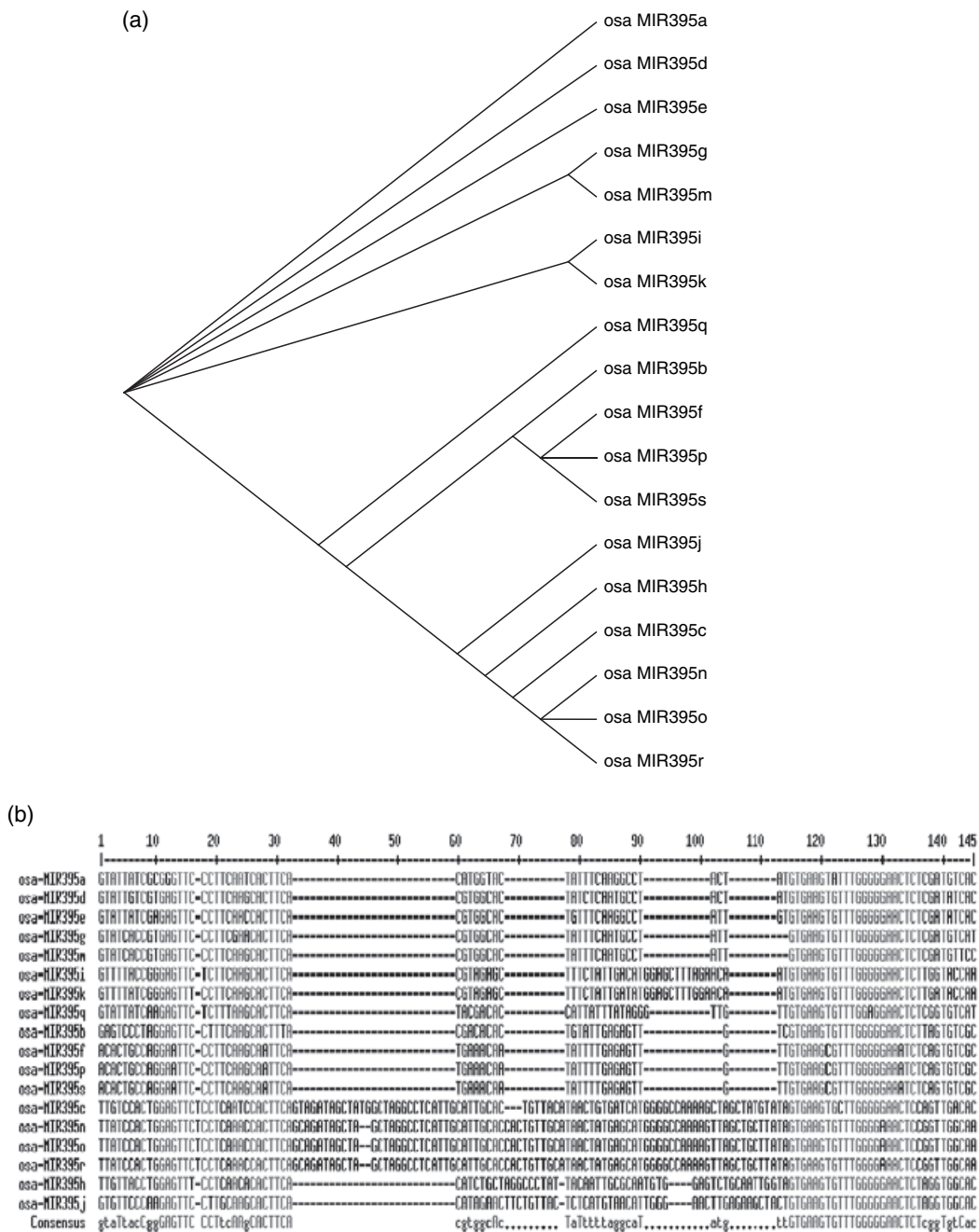


Figure 10. Polygenetic tree (a) and multiple sequence alignment (b) of the 18 members of rice miRNA 395.

our results. Previous studies demonstrated that all miRNAs regulate gene expression by binding to target mRNA sequences at a perfect or near-perfect complementary site, indicating that plant miRNA targets can be recognized using a simple homology search. In this study, we used the same procedure described above for predicting miRNA homologs to predict potential miRNA targets in a protein-coding gene database instead of a genomic database. We tested miRNA 172 against the GenBank protein database with BLASTn (Altschul *et al.*, 1997) using the same parameters described in the miRNA search. The number of allowed mismatches at comple-

mentary sites between miRNA sequences and potential mRNA targets was three or fewer, and no gaps were allowed at the complementary sites.

Minimizing false positives

Existence of false positives can be a problem in these types of studies. To reduce the number of false positives, we combined multiple methods in this study, similar to that described by Bonnet

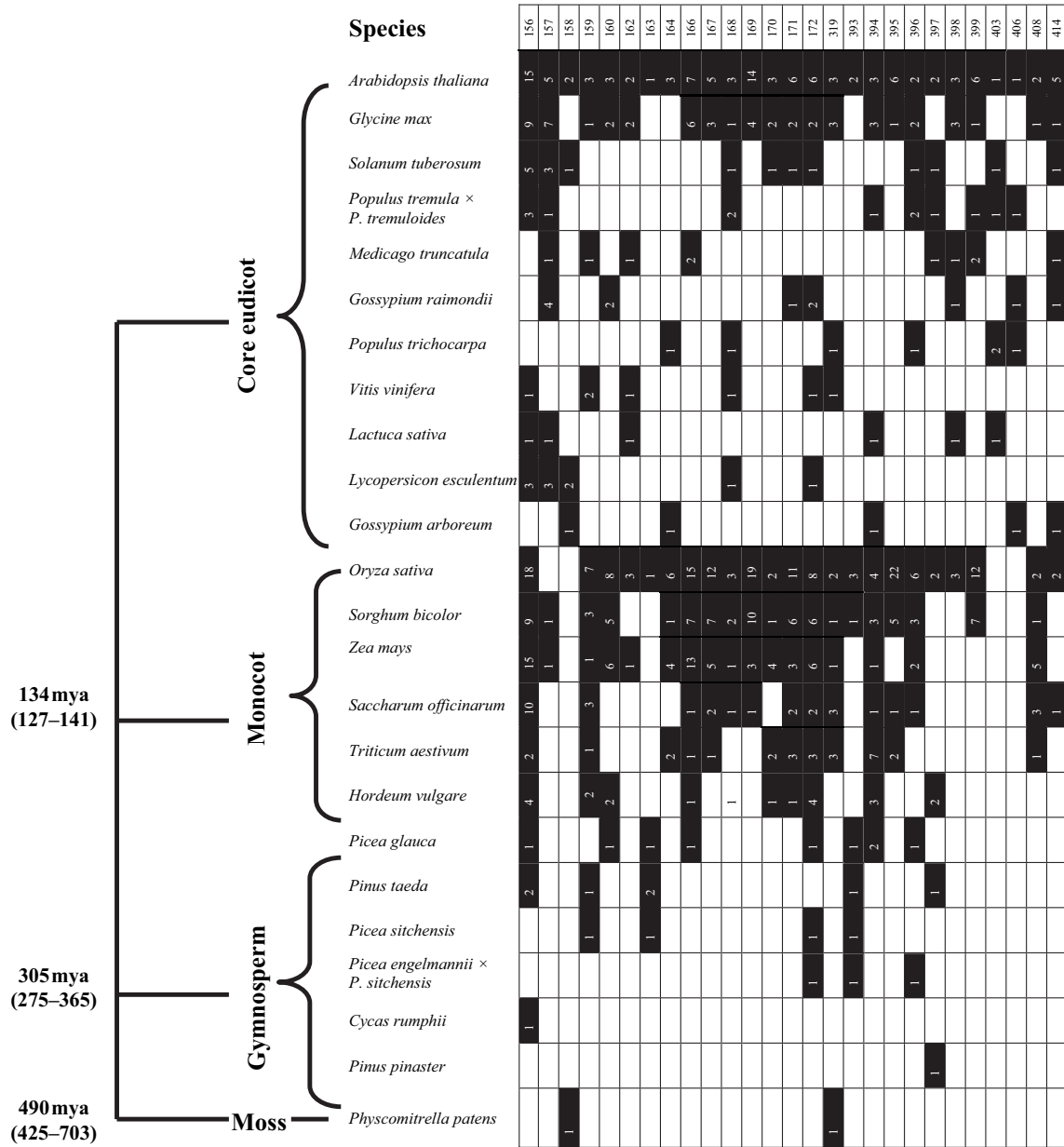


Figure 11. Majority of miRNA families were found in various plant species representing major plant subgroups. This indicates that miRNAs diverged at a very early stage of plant phylogeny about 425 million years ago (mya). The top number on the left at the major nodes indicates the most reliable estimate, while the numbers below in parentheses are the range of estimates that have been published for that node (Sanderson *et al.*, 2004).

et al. (2004a). First, the number of EST hits was dramatically reduced after comparing EST sequences to previously known *Arabidopsis thaliana* miRNA sequences. Based on the fact that plant miRNAs are highly conserved and only a few nucleotides change between plant species, ESTs with only 0–3 mismatched nucleotides with previously known *Arabidopsis thaliana* miRNAs were considered as potential miRNA candidates. Second, potential ESTs were reduced about 50% by considering the secondary structure parameters based on previous reports (Bartel, 2004; Bartel and Bartel, 2003; Bonnet *et al.*, 2004a; Reinhart *et al.*, 2002; Sunkar and Zhu, 2004). Third, the number of miRNAs was reduced by considering potential miRNA

target genes in the DNA database. Fourth, repeated ESTs were removed by comparing similar EST sequences. Fifth, to avoid a miscount of the total EST sequences, ESTs with *n*-2, *n*-3, *n*-4 and *n*-5 mismatched nucleotides with previously known miRNAs were chosen and paired with known miRNA sequences. Sixth, the matched ESTs were blasted against a database of known proteins to kick out the potential ESTs that actually code a protein rather than a structural RNA. After these six steps and other combined strategies, the number of EST-predicted miRNAs was dramatically reduced. Finally, to further confirm our conclusions, we also searched the potential targets of EST-predicted miRNAs as another type of

evidence for the real existence of those miRNAs in a given plant species.

References

- Achard, P., Herr, A., Baulcombe, D.C. and Harberd, N.P. (2004) Modulation of floral development by a gibberellin-regulated microRNA. *Development*, **131**, 3357–3365.
- Adai, A., Johnson, C., Mlotshwa, S., Archer-Evans, S., Manocha, V., Vance, V. and Sundaresan, V. (2005) Computational prediction of miRNAs in *Arabidopsis thaliana*. *Genome Res.* **15**, 78–91.
- Adams, M.D., Kelley, J.M., Gocayne, J.D. et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
- Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M.J., Tuschl, T. and Margalit, H. (2005) Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res.* **33**, 2697–2706.
- Ambros, V. (2001) microRNAs: tiny regulators with great potential. *Cell*, **107**, 823–826.
- Ambros, V., Bartel, B., Bartel, D.P. et al. (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.
- Aukerman, M.J. and Sakai, H. (2003) Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell*, **15**, 2730–2741.
- Axtell, M.J. and Bartel, D.P. (2005) Antiquity of microRNAs and their targets in land plants. *Plant Cell*, **17**, 1658–1673.
- Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
- Bartel, B. and Bartel, D.P. (2003) MicroRNAs: at the root of plant development? *Plant Physiol.* **132**, 709–717.
- Baskerville, S. and Bartel, D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**, 241–247.
- Bonnet, E., Wuyts, J., Rouze, P. and Van de Peer, Y. (2004a) Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc. Natl Acad. Sci. USA*, **101**, 11511–11516.
- Bonnet, E., Wuyts, J., Rouze, P. and Van de Peer, Y. (2004b) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.
- Carrington, J.C. and Ambros, V. (2003) Role of microRNAs in plant and animal development. *Science*, **301**, 336–338.
- Chen, X. (2004) A microRNA as a translational repressor of APETALA2 in *Arabidopsis* flower development. *Science*, **303**, 2022–2025.
- Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881–10890.
- Eckardt, N.A. (2005) MicroRNAs regulate auxin homeostasis and plant development. *Plant Cell*, **17**, 1335–1338.
- Floyd, S.K. and Bowman, J.L. (2004) Gene regulation: ancient microRNA target sequences in plants. *Nature*, **428**, 485–486.
- Griffiths-Jones, S. (2004) The microRNA Registry. *Nucleic Acids Res.* **32**, D109–D111.
- Guo, H.S., Xie, Q., Fei, J.F. and Chua, N.H. (2005) MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for *Arabidopsis* lateral root development. *Plant Cell*, **17**, 1376–1386.
- Hunter, C. and Poethig, R.S. (2003) miSSING LINKS: miRNAs and plant development. *Curr. Opin. Genet. Dev.* **13**, 372–378.
- Inukai, Y., Sakamoto, T., Ueguchi-Tanaka, M., Shibata, Y., Gomi, K., Umemura, I., Hasegawa, Y., Ashikari, M., Kitano, H. and Matsuo, M. (2005) Crown rootless1, which is essential for crown root formation in rice, is a target of an AUXIN RESPONSE FACTOR in auxin signaling. *Plant Cell*, **17**, 1387–1396.
- Jack, T. (2004) Molecular and genetic mechanisms of floral control. *Plant Cell*, **16**, S1–S17.
- Jones-Rhoades, M.W. and Bartel, D.P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell*, **14**, 787–799.
- Kasschau, K.D., Xie, Z., Allen, E., Llave, C., Chapman, E.J., Krizan, K.A. and Carrington, J.C. (2003) P1/HC-Pro, a viral suppressor of RNA silencing, interferes with *Arabidopsis* development and miRNA function. *Dev. Cell*, **4**, 205–217.
- Kidner, C.A. and Martienssen, R.A. (2005) The developmental role of microRNA in plants. *Curr. Opin. Plant Biol.* **8**, 38–44.
- Lai, E.C. (2003) microRNAs: runts of the genome assert themselves. *Curr. Biol.* **13**, R925–R936.
- Lai, E.C., Tomancak, P., Williams, R.W. and Rubin, G.M. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4**, R42.
- Laufs, P., Peaucelle, A., Morin, H. and Traas, J. (2004) MicroRNA regulation of the CUC genes is required for boundary size control in *Arabidopsis* meristems. *Development*, **131**, 4311–4322.
- Lauter, N., Kampani, A., Carlson, S., Goebel, M. and Moose, S.P. (2005) microRNA172 down-regulates *glossy15* to promote vegetative phase change in maize. *Proc. Natl Acad. Sci. USA*, **102**, 9412–9417.
- Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, **75**, 843–854.
- Lee, Y., Ahn, C., Han, J. et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.
- Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B. and Bartel, D.P. (2003a) Vertebrate microRNA genes. *Science*, **299**, 1540.
- Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., Bartel, D.P. (2003b) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17**, 991–1008.
- Llave, C., Xie, Z., Kasschau, K.D. and Carrington, J.C. (2002) Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science*, **297**, 2053–2056.
- Lohmann, J.U. and Weigel, D. (2002) Building beauty: the genetic control of floral patterning. *Dev. Cell*, **2**, 135–142.
- Mallory, A.C., Bartel, D.P. and Bartel, B. (2005) MicroRNA-directed regulation of *Arabidopsis* AUXIN RESPONSE FACTOR17 is essential for proper development and modulates expression of early auxin response genes. *Plant Cell*, **17**, 1360–1375.
- Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940.
- Matukumalli, L.K., Grefenstette, J.J., Sonstegard, T.S. and Van Tassell, C.P. (2004) EST-PAGE – managing and analyzing EST data. *Bioinformatics*, **20**, 286–288.
- Palatnik, J.F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J.C. and Weigel, D. (2003) Control of leaf morphogenesis by microRNAs. *Nature*, **425**, 257–263.
- Park, W., Li, J., Song, R., Messing, J. and Chen, X. (2002) CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12**, 1484–1495.
- Pasquinelli, A.E., Reinhart, B.J., Slack, F. et al. (2000). Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature*, **408**, 86–89.

- Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B. and Bartel, D.P.** (2002) MicroRNAs in plants. *Genes Dev.* **16**, 1616–1626.
- Sanderson, M.J., Thorne, J.L., Wikstrom, N. and Bremer, K.** (2004) Molecular evidence on plant divergence times. *Am. J. Bot.* **91**, 1656–1665.
- Seitz, H., Royo, H., Bortolin, M.L., Lin, S.P., Ferguson-Smith, A.C. and Cavaille, J.** (2004) A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. *Genome Res.* **14**, 1741–1748.
- Sunkar, R. and Zhu, J.K.** (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell*, **16**, 2001–2019.
- Swofford, D.L.** (2000) *PAUP*: Phylogenetic Analysis Using Parsimony (*and other Methods), Version 4*. Sunderland, UK: Sinauer Associates.
- Tanzer, A. and Stadler, P.F.** (2004) Molecular evolution of a microRNA cluster. *J. Mol. Biol.* **339**, 327–335.
- Wang, X.J., Reyes, J.L., Chua, N.H. and Gaasterland, T.** (2004) Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol.* **5**, R65.
- Weber, M.J.** (2005) New human and mouse microRNA genes found by homology search. *FEBS J.* **272**, 59–73.
- Wightman, B., Ha, I. and Ruvkun, G.** (1993) Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell*, **75**, 855–862.
- Zhang, B.H., Pan, X.P., Wang, Q.L., Cobb, G.P. and Anderson, T.A.** (2005) Identification and characterization of new plant microRNAs using EST analysis. *Cell Res.* **15**, 336–360.
- Zhang, B.H., Pan, X.P., Cobb, G.P. and Anderson, T.A.** (2006a). Plant microRNA: a small regulatory molecule with big impact. *Dev. Biol.* **289**, 3–16.
- Zhang, B.H., Pan, X.P., Cox, S.B., Cobb, G.P. and Anderson, T.A.** (2006b) Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.* **63**, 246–254.
- Zuker, M.** (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415.