Evolution of *Arabidopsis* microRNA families through duplication events

Christopher Maher, Lincoln Stein and Doreen Ware

Genome Res. 2006 16: 510-519
Access the most recent version at doi:10.1101/gr.4680506

Letter

Evolution of *Arabidopsis* microRNA families through duplication events

Christopher Maher,[1,2,4] Lincoln Stein,[1] and Doreen Ware[1,3]

[1]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA;* [2]*Department of Biomedical Engineering, Stony Brook University, Stony Brook, New York 11794, USA;* [3]*United States Department of Agriculture–Agricultural Research Service (USDA–ARS) North Atlantic Area (NAA) Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, New York 15853, USA*

Recently there has been a great interest in the identification of microRNAs and their targets as well as understanding the spatial and temporal regulation of microRNA genes. To understand how microRNA genes evolve, we looked at several rapidly evolving families in *Arabidopsis thaliana*, and found that they arose from a process of genome-wide duplication, tandem duplication, and segmental duplication followed by dispersal and diversification, similar to the processes that drive the evolution of protein gene families. Using multiple expression data sets to examine the transcription patterns of different members of the microRNA families, we find the sequence diversification of duplicated microRNA genes to be accompanied by a change in spatial and temporal expression patterns, suggesting that duplicated copies acquire new functionality as they evolve.

[Supplemental material is available online at www.genome.org.]

It has been suggested that microRNAs, or miRNAs, play a central role in regulating basic developmental processes, such as meristem cell identity, organ polarity, and timing of developmental events, by interfering with the expression of targeted messenger RNAs (mRNAs) (Emery et al. 2003; Palatnik et al. 2003; Bartel 2004). Understanding the role of miRNAs could help answer fundamental biological questions while also enhancing the ability to precisely engineer plants for improved crop yields, increased resistance to disease, and adaptation to environmental extremes.

miRNAs are a class of small single-stranded non-coding RNAs that range in length from roughly 20 to 24 nucleotides (nt) (Bartel and Bartel 2003; Bartel 2004). The biogenesis of miRNAs differs between plants and animals. Within plants, it is believed that polymerase II transcribes miRNAs into a primary miRNA transcript (pri-miRNA). In the nucleus, a ribonuclease III-like nuclease, DICER-LIKE 1 (DCL1) (Papp et al. 2003), then processes the pri-miRNA, potentially with the assistance of one or more unknown enzymes. This process yields the precursor miRNA (pre-miRNA) and ultimately the mature miRNA:miRNA* duplex (Bartel 2004). The mature miRNA duplex is exported to the cytoplasm, where it is unwound and incorporated into the RISC complex (Bartel 2004). The miRNA then guides the complex to its specific protein-coding gene target mRNA, partially or completely silencing the transcript by either degrading it or by inhibiting its translation into a protein (Llave et al. 2002).

Plant miRNAs can be grouped into distinct families of one or more precursors. Each precursor within the family produces similar, if not identical, mature miRNA products. Within a family, the greatest sequence conservation occurs in the stem that becomes the mature miRNA product, followed by the stem that opposes the mature miRNA in the precursor. Within both plants and animals, the unpaired loop regions are the most variable parts of the precursor despite the characteristically smaller loop lengths found in animal hairpins (Lai et al. 2003; Maher et al. 2004). High levels of sequence similarity among loop regions of *Arabidopsis* precursors appear only in tandemly duplicated precursors (Maher et al. 2004). In most cases, there is no obvious sequence similarity among the loop regions of members of the same miRNA family.

Direct evidence pertaining to the mechanism of miRNA transcription has only recently been published (Lee et al. 2004; Xie et al. 2005). Currently, the majority of plant miRNAs reside within intergenic regions or in the opposite strand of annotated genes. miRNAs, like mRNAs, are transcribed by polymerase II. We therefore expect miRNA sequences to be found in collections of Pol II-transcribed RNAs, such as Massively Parallel Signature Sequencing (MPSS) collections.

In *Arabidopsis,* protein-coding gene families arise by a process of gene duplication and diversification (The *Arabidopsis* Genome Initiative 2000; Prince and Pickett 2002; Cannon et al. 2004). The processes driving gene duplication are whole-genome duplication (polyploidization), duplications of subchromosomal-length regions known as segmental duplications, and local duplications that involve one or two genes known as tandem duplications (Bowers et al. 2003; Lawton-Rauh 2003; Blanc and Wolfe 2004b). Gene- and chromosomal-level rearrangements increase the difficulty of numbering and dating polyploidy events (Lawton-Rauh 2003; Blanc and Wolfe 2004a; Adams and Wendel 2005).

The goal of this study was to ask whether this model of protein-coding gene family evolution applies to the miRNA gene families as well, and, if so, whether there exists an association between the evolution of miRNA genes and changes in expression patterns that might indicate diversification of function.

Results

The haploid genome of *Arabidopsis* consists of five chromosomes containing many internally duplicated regions. To begin this work, we obtained all 92 *Arabidopsis* miRNA precursor gene sequences and coordinates from the miRNA Registry (http://microrna.sanger.ac.uk/) (Ambros et al. 2003; Griffiths-Jones

[4]Corresponding author.
E-mail maher@cshl.edu; fax (516) 367-6851.
Article published online ahead of print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4680506.

16:510–519 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

2004). The miRNA genes were grouped into 26 families based on the similarity of the mature miRNA product (Ambros et al. 2003). Of the 26 families, 22 (84.6%) contain more than one miRNA gene and six families (25%) contain five or more miRNA genes (Supplemental Table 1). Given the large number of miRNA families with multiple genes, it is reasonable to hypothesize that they have undergone a history of expansion events similar to those that underlie the amplification and diversification of families of protein-coding genes. Therefore, we expect to see different members from the same miRNA family residing within duplicated regions of the genome.

Tandem duplications

We first identified apparent tandem duplications among the miRNA gene families. We did so by looking for contiguous miRNAs in the same intergenic region, or in neighboring intergenic regions, and found 23 genes from six gene families that met these criteria. The longest run of miRNA genes arising from an apparent tandem duplication was six, while the remainder occurred in arrays of two or three miRNA genes. Of the 23 tandemly duplicated miRNAs, if each miRNA is paired with the nearest downstream tandemly duplicated miRNA, two-thirds are on the same strand and the average distance between tandemly duplicated miRNAs is 1987 nt (data not shown).

Large duplication events

We next wished to test the hypothesis that large-scale duplication events play a role in the evolution of miRNA gene families. We reasoned that, if this were the case, then the protein-coding genes flanking members of the same miRNA family would be more similar to each other than protein-coding genes flanking randomly selected genes, because the protein-coding genes would also be involved in the duplications. The alternative hypothesis is that miRNAs are not evolving through duplication events, but rather via random translocations and insertional events. To identify large duplication events, we chose to align protein-coding genes rather than non-coding nucleotide sequence because of the low level of nucleotide sequence conservation among non-coding regions in *Arabidopsis* duplicated regions (Vision et al. 2000). We therefore consider miRNAs to originate from a duplication event if they reside within a region of conserved protein-coding genes. Two chromosomal regions, containing one or more miRNAs, were classified as residing within such a duplicated block if one or more of the 10 upstream or 10 downstream protein-coding genes flanking the miRNA were found to have a best non-self match to a protein-coding gene flanking another miRNA according to BLASTP (*E*-value < 0.001). Only the best match was used for this analysis so that tandemly duplicated genes did not enrich the number of conserved genes flanking a miRNA. In addition, using the best match selects for paralogs that are more likely to be recently duplicated from one another over less conserved genes from the same family. As a control, we generated a simulated data set in which we selected random genomic locations and aligned their flanking protein-coding genes.

Our approach excluded miRNA families containing a single gene and therefore leaves us with 88 miRNA precursors from 22 distinct miRNA families. Since we are aligning the flanking protein-coding genes for a miRNA, tandemly duplicated miRNAs were counted only once. Therefore, our 88 miRNA precursors were located within 73 chromosomal regions.

To characterize the pattern of miRNA duplication, we compared the rates of duplicated blocks surrounding miRNAs within the same family (intrafamily), between families (interfamily), and randomly selected locations (Fig. 1). In our analysis we found that there are 26 duplicated chromosomal regions containing miRNAs from the same family that have conservation between their flanking protein-coding genes out of the 116 total possible miRNA pairs (22.42%) as opposed to 1.3% of interfamily miRNA pairs and 1.94% of randomly selected genomic locations. Together, these data suggest that large-scale duplication plays a major role in miRNA evolution and are inconsistent with the random insertion hypothesis. Our procedure may misclassify duplicated blocks at a rate of ~2%.

While the randomized set represents the upper bound of our false-positive rate, we also observed that the randomized and interfamily duplicated blocks tend to have fewer conserved protein-coding regions than the intrafamily duplicated blocks. In fact, interfamily duplicated blocks all occur with three or fewer conserved flanking genes, with the exception of the *miR169a–miR158b* pair, which has 12 conserved flanking genes. Therefore, we believe our classification system is more likely to fail when applied to duplicated blocks having three or fewer conserved flanking protein-coding genes. Almost half of the putative intrafamily duplicated blocks that we have identified have at least three conserved flanking genes. We therefore defined all predicted duplicated blocks as our "loose" set and the duplicated blocks containing four or more conserved flanking protein-coding genes as our "strict" set.

While our previous methodology analyzed 10 upstream and downstream protein-coding genes in order to identify duplicated blocks, these duplicated regions can span much larger regions, which we will refer to as extended duplicated blocks. To enable a more detailed analysis of miRNA families, we wanted to provide a broad overview of each duplicated region and therefore extracted 200 protein-coding genes flanking each miRNA. We then plotted these protein-coding genes surrounding the miRNAs to highlight our previously identified duplicated blocks, but in addition show the varying degrees of chromosomal rearrangements, if any, within the extended duplicated block. This enables us to establish relationships between miRNAs that are more closely related to one another within a particular family. In addition, we incorporate expression data to further support the diversification of miRNAs.

Table 1 summarizes the number of segmental and tandem duplications for each miRNA family according to our definitions. It would appear that 18 of the 22 families (81.8%) arise from either a segmental or tandem duplication, or a combination of the two processes. Of these 18 families, six were involved in tandem duplications, and 17 were involved in segmental duplications. In total, 23 (26.1%) miRNAs are involved in tandem duplications, while 51 miRNAs (57.9%) are involved in large-scale duplication events. A more conservative estimate of segmental duplications, which would discard all miRNAs that have three or fewer conserved flanking protein-coding genes, predicts that 32 miRNAs (36.3%) would be involved in duplicated blocks. This suggests that miRNA genes are evolving by segmental duplications and tandem duplications, just as protein-coding genes have evolved.

Dating duplication events

Under the assumption that synonymous silent substitutions per site (K_s) occur with a constant rate over time, we can use the

A

Duplicated miRNA 1	Dulicated miRNA 2	Conserved Flanking Protein-coding Genes
ath-miR156c	*ath-miR156a*	8
ath-miR394a	*ath-miR394b*	7
ath-miR166a	*ath-miR166b*	6
ath-miR399def	*ath-miR399a*	6
ath-miR159b	*ath-miR159a*	5
ath-miR160b	*ath-miR160c*	5
ath-miR395abc	*ath-miR395def*	4
ath-miR162b	*ath-miR162a*	4
ath-miR165b	*ath-miR165a*	4
ath-miR166g	*ath-miR166cd*	4
ath-miR169f	*ath-miR169de*	4
ath-miR393a	*ath-miR393b*	4
ath-miR157c	*ath-miR157d*	3
ath-miR167a	*ath-miR167b*	3
ath-miR167b	*ath-miR167c*	2
ath-miR156e	*ath-miR156f*	2
ath-miR171b	*ath-miR171c*	2
ath-miR156a	*ath-MI156d*	1
ath-miR156d	*ath-miR156c*	1
ath-miR156g	*ath-miR156e*	1
ath-miR156f	*ath-miR156h*	1
ath-miR164a	*ath-miR164c*	1
ath-miR167c	*ath-miR167a*	1
ath-miR172c	*ath-miR172b*	1
ath-miR172d	*ath-miR172c*	1
ath-miR319a	*ath-miR319b*	1

B

Conserved Genes	Observed Intra-miRNA Duplications	Percent of total Intra-miRNA Duplication Events	Observed Inter-miRNA Duplications	Percent of Total Inter-miRNA Duplication Events	Observed Random Duplications	Percent of Total Random Duplication Events
1	9	7.76	31	1.04	12	1.2
2	3	2.59	3	0.1	2	0.2
3	2	1.73	4	0.14	1	0.1
4	6	5.18	0	0	1	0.1
5	2	1.73	0	0	1	0.1
6	2	1.73	0	0	0	0
7	1	0.87	0	0	0	0
8	1	0.87	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	0
12	0	0	1	0.04	0	0
Total	26	22.42	39	1.3	17	1.7

C

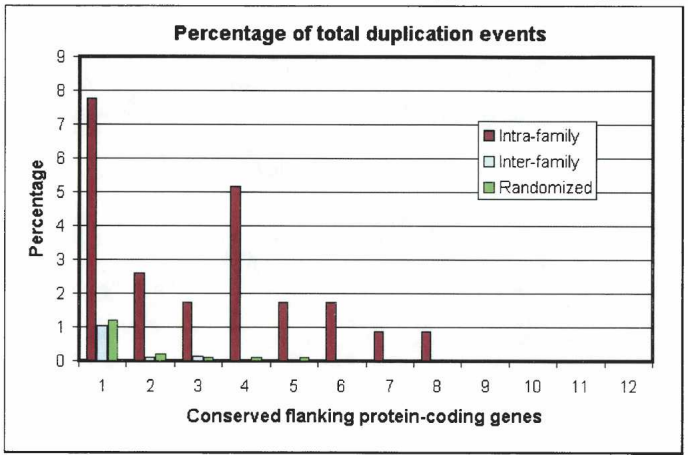


Figure 1. Percentage of intrafamily, interfamily, and randomly duplicated blocks. The total number of duplication events is the sum of all possible miRNA pairs within the set. Therefore, the percentage equals the total number of duplication events observed compared to the total number of possible duplication events. (*A*) Duplicated miRNAs from within the same family with the number of conserved protein-coding genes flanking the miRNAs. (*B*) Percentages of observed duplicated blocks against the total number of potential duplicated blocks and the number of flanking genes that are conserved within each block. (*C*) Plot comparing the percentage of observed duplication events against the total number of potential duplications for interfamily miRNAs, intrafamily miRNAs, and the randomized simulation.

conserved flanking protein-coding genes to estimate the dates of the large-scale duplication events. For this analysis, we used duplicated blocks in our strict set only. Each pair of proteins in the duplicated block was aligned at the amino acid level, and then codons from gapless aligned regions were used to calculate K_s values using codeml (Yang 1997). We discarded any K_s values >2.0 because of the risk of saturation (Blanc and Wolfe 2004b). The approximate date of the duplication event was then calculated using the mean K_s and an estimated rate of silent-site substitutions of 1.5×10^{-8} substitutions/synonymous site/year (Koch et al. 2000; Blanc and Wolfe 2004b). Table 2 shows the mean K_s values for each duplication event and the estimated date. We conclude that the large-scale duplication events involving miRNAs have all occurred within the last 28–39 million years (Myr). Given that traces of duplication events erode with time, we believe our approach may be limited to duplication events that have occurred within the last 39 Myr. Thus, miRNAs lacking conserved flanking protein-coding genes, which nevertheless maintain sequence conservation across both stems of their precursor, may have evolved prior to the events we have detected.

Relationship of miRNAs and their targets

For multigene miRNA families that target multiple mRNAs, with similar or identical target sites, we were interested to see whether there was a correlation in the physical locations of known miRNAs and their targets. If so, miRNAs in close proximity to their respective target mRNAs could be indicative of a regulatory relationship. Previous studies have identified potential miRNA targets based on a predetermined set of rules for base-pairing between a miRNA and its target mRNA (Rhoades et al. 2002; Jones-Rhoades and Bartel 2004; Schwab et al. 2005). Using these predicted targets, we observed that precursors from within the same and different families are scattered physically throughout the genome and that there is no apparent correlation between miRNA genes and their protein-coding targets (Supplemental Fig. 1).

Expansion of miRNA families in conjunction with expression data

When protein-coding genes duplicate and diverge, they can lose function (become pseudogenes), maintain their current function (redundant function), acquire new functions (neofunctionalization), or take on more specialized functions (subfunctionalization). We next asked whether the same processes apply to the miRNA genes. To answer this question, we obtained spatial- and temporal-specific expression pattern data from Massively Parallel Signature Sequencing (MPSS) collections (Meyers et al. 2004a).

Table 1. Duplication events of miRNAs in multigene families

miRNA family	Loci	Loci in tandem duplication(s)	Loci in large-scale duplication(s) loose definition	Loci in large-scale duplication(s) strict definition	Target mRNAs
miR156	8	—	7	2	Squamosa-promoter binding protein (SPB)-like proteins
miR157	4	2	2	0	Squamosa-promoter binding protein (SPB)-like proteins
miR158	2	—	—	—	PPR repeat protein
miR159	3	—	2	2	MYB proteins/TCP transcription factors
miR160	3	—	2	2	Auxin response factors (ARF transcription factors)
miR162	2	—	2	2	DICER-LIKE 1
miR164	3	—	2	—	NAC domain proteins
miR165	2	—	2	2	HD-Zip transcription factors
miR166	7	2	5	5	HD-Zip transcription factors
miR167	4	—	3	—	Auxin response factor
miR168	2	—	—	—	ARGONAUTE
miR169	14	8	3	3	CCAAT-binding factor (CBF)-HAP2-like proteins
miR171	3	—	2	—	GRAS domain proteins (SCARECROW-like)
miR172	5	—	3	—	APETALA2-like transcription factors
miR319	3	—	2	—	MYB and TCP transcription factors
miR393	2	—	2	2	F-box proteins and bHLH transcription factors
miR394	2	—	2	2	F-box proteins
miR395	6	6	6	6	ATP sulphurylases
miR396	2	—	—	—	Growth regulating factor (GRF) transcription factors, rhodenase-like proteins, and kinesin-like protein B
miR397	2	—	—	—	Laccases and β-6 tubulin
miR398	3	2	—	—	Copper superoxide dismutases and cytochrome *C* oxidase subunit V
miR399	6	3	4	4	Phosphatase transporter
	88	23	51	32	

This table indicates the number of loci within a family found to be tandemly duplicated or within a duplicated block, along with the target mRNA. The number of segmental duplications is shown under both a loose and a strict definition. The loose definition shows all possible miRNAs that reside in duplicated blocks, while the strict definition shows the number of miRNAs in duplicated blocks with four or more conserved flanking genes.

MPSS is a large-scale expression resource capturing transcript expression levels within 17 different libraries. The MPSS signatures are derived from the 3′-end of the mRNA molecule (Meyers et al. 2004a). Therefore, mapping the signature relative to the miRNA should show a higher density of signatures downstream of the miRNA (Meyers et al. 2004b). It is a possibility that some of the miRNAs are alternatively spliced, given the close proximity of multiple signatures downstream of the miRNA, but for our purposes we were interested in the minimum distance of downstream signatures. Therefore, for each miRNA, we record only the first occurring significantly expressed signature downstream of the miRNA yet upstream of the adjacent protein-coding gene, as shown in Table 3. We observed a greater density of expressed class 4 (intergenic) signatures located slightly downstream of the known miRNAs within the first 400 nt (data not shown).

We analyzed the 92 miRNAs from 26 different families and merged the 19 tandemly duplicated miRNAs that reside within the same intergenic region since it is not known whether they are expressed as one large transcriptional unit or as two separate primary transcripts. Overall, 32 of the 92 miRNAs (34.8%) have an associated class 4 signature that is expressed at significant levels, as shown in Table 3, assuming the two tandemly duplicated miRNAs are polycistronic. The average expression level is 26 transcripts per million (TPM), with a range of 4–173. We then correlated the tissue distribution of the MPSS signatures associated with each known miRNA (Supplemental Table 2).

For those miRNAs that did not have an MPSS signature, it is possible that their expression patterns are specific to tissues not sampled by the MPSS libraries. This is consistent with a recent analysis of *Arabidopsis* miRNA gene expression, in which 47 out of 99 (47.4%) miRNAs failed to produce a detectable signal using 5′-RACE or 3′-RACE (Xie et al. 2005). Of the 52 miRNAs detected by RACE, 25 (48.1%) miRNAs have an associated MPSS signature (Table 3). Of 47 miRNAs not detected by RACE, nine (14.9%) miRNAs have an associated MPSS signature. Overall, the miRNAs for which we failed to find MPSS signatures were more likely to be undetectable by RACE.

In the following sections, we describe specific examples of how the miR156, miR159, and miR166 families seem to evolve and take on new functionality through duplication events.

Table 2. Estimation of the absolute date for large-scale duplication events

Duplicated pair	*n*	Mean K_s	SD K_s	Minimum K_s	Maximum K_s	Date (Myr)
156a/c	7	0.979	0.19	0.7	1.23	32.64
159a/b	5	0.8	0.17	0.6	0.94	26.67
160b/c	5	1.16	0.16	1.01	1.37	38.67
162a/b	4	0.98	0.33	0.78	1.35	32.67
165a/b	4	1.06	0.09	0.94	1.14	35.34
166a/b	6	1.01	0.28	0.65	1.41	33.67
166cd/g	4	1.02	0.4	0.66	1.56	34
169de/f	4	0.81	0.13	0.7	0.99	27
393a/b	4	1.13	0.43	0.83	1.74	37.67
394a/b	7	1.17	0.32	0.68	1.63	39
395abc/def	4	1.01	0.04	0.98	1.06	33.67
399a/def	5	0.84	0.2	0.64	1.08	28
						Average: 33.25

For each duplicated region containing miRNAs, we indicate the number of protein-coding genes, *n*, used for the K_s estimation. Only duplication events containing four or more conserved protein-coding genes were used to calculate the duplication event date. The events range from 28 to 39 million years ago (Mya), with the average date occurring around 33.5 Myr.

Table 3. Class 4, intergenic MPSS signatures for known miRNAs

miRNA	Neighboring genes		MPS signature	Nucleotides downstream of foldback structure	Transcripts per million (TPM)	5′-RACE	3′-RACE	Small RNAs in ASRP
miR156a	At2g25090	At2g25100	GATCTCTTTGGCCTGTC	11	6	Yes	NT	2
miR156b	At4g30970	At4g30980	GATCGCTTCTTATCATC	50	6	No	No	2
miR156d	At5g10940	At5g10950	GATCGAATAAGGGATGT	806	127	No	NT	3
miR156f	At5g26140	At5g26150	GATCGCCACACCCTCCC	152	4	Yes	NT	3
miR157a	At1g66780	At1g66790	GATCCGACTGAAAGGAT	240	74	No	Yes	2
miR157b	At1g66790	At1g66800	GATCATTGTCCAGATTC	1013	4	No	Yes	2
miR157c	At3g18215	At3g18220	GATCTTTGGATTCGACC	719	19	Yes	NT	2
miR158b	At1g55590	At1g55600	GATCTTGCTCTAAACTT	226	4	No	No	1
miR159a	At1g73680	At1g73690	GATCCTTGGTTCTTTGG	226	65	Yes	NT	4
miR159b	At1g18070	At1g18080	GATCTTGAGTAGGATTT	95	173	Yes	NT	3
miR164b	At5g01740	At5g01750	GATCACTATTAGTAATC	1302	7	Yes	NT	1
miR165a	At1g01180	At1g01190	GATCCGTCTATGCTTTT	139	15	Yes	NT	1
miR166a	At2g46680	At2g46690	GATCTCTTACCTTACTT	73	22	Yes	NT	4
miR166b	At3g61890	At3g61900	GATCTTCTGAGTTTCAG	167	37	Yes	NT	2
miR166c	At5g08710	At5g08720	GATCCTGAAGTGAGAGC	2057	14	Yes	NT	2
miR166d	At5g08710	At5g08720	GATCCTGAAGTGAGAGC	96	14	Yes	NT	2
miR166f	At5g43600	At5g43610	GATCACCTAATTCTCTA	40	21	No	Yes	1
miR167b	At3g63370	At3g63380	GATCTATCATAGGTGCA	163	9	Yes	NT	5
miR168a	At4g19390	At4g19400	GATCTGGAAGATTTCTA	24	24	No	NT	1
miR169a	At3g13400	At3g13410	GATCTTGATGAATTTCA	281	16	Yes	NT	7
miR169m	At3g26810	At3g26820	GATCAATTCTTCAGAGA	366	4	No	NT	6
miR169n	At3g26810	At3g26820	GATCAATTCTTCAGAGA	738	4	Yes	NT	5
miR170	At5g66040	At5g66050	GATCGGATGCTCCTTTC	186	8	Yes	NT	1
miR171a	At3g51380	At3g51390	GATCTTGTCTTCTTTTG	881	45	Yes	NT	2
miR171b	At1g11730	At1g11740	GATCGGTAGCCTTAGAG	116	21	Yes	NT	1
miR172a	At2g28050	At2g28056	GATCTGACAAAATGAGA	1544	11	Yes	NT	1
miR172b	At5g04270	At5g04280	GATCGGCCAGTTCGGTC	392	41	Yes	NT	1
miR393b	At3g55730	At3g55740	GATCCAGTCATATCAAC	70	11	No	No	1
miR394a	At1g20370	At1g20380	GATCAAGGAATAGGTGA	1133	6	Yes	NT	1
miR395e	At1g69790	At1g69800	GATCCGACATGTTTAAA	240	5	Yes	NT	1
miR399a	At1g29260	At1g29270	GATCTAAAAGTTCACGG	991	10	No	No	1
miR399c	At5g62160	At5g62165	GATCTAAAGTCTAAAAA	224	7	Yes	NT	1

For each miRNA with an associated MPSS signature downstream, the neighboring protein-coding genes, the 17-nt MPSS signature, nucleotide distance downstream of the precursor 3′-end, and expression level are shown. The expression level is normalized to show how many transcripts, containing the signature, occurred for every million different transcripts captured within the library. Tandemly duplicated miRNAs were kept in this table to show the specific small RNA, 5′-RACE, and 3′-RACE results for each gene despite having the same MPSS signature. 5′-RACE and 3′-RACE values are indicated as Yes, No, or NT (Not Tested) (Xie et al. 2005). The small RNAs in ASRP represent the number of clones related to that particular miRNA gene (Gustafson et al. 2005).

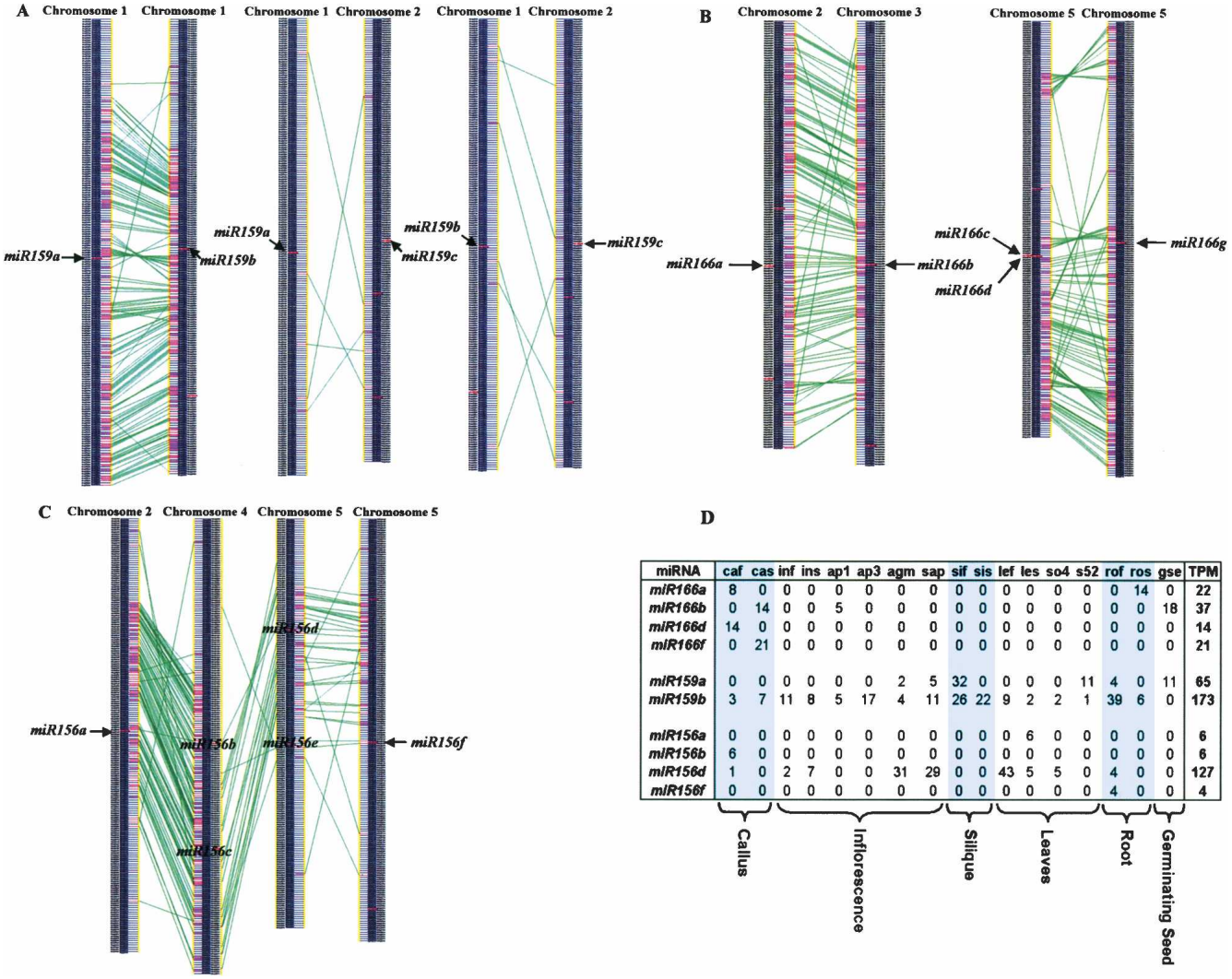
miR159 family evolution

The three precursors within the miR159 family target mRNAs coding for MYB proteins, which are known to bind to the promoter of the floral meristem identity gene LEAFY and have varying degrees of conservation in their surrounding regions (Reinhart et al. 2002; Rhoades et al. 2002; Achard et al. 2004). Using the extended duplicated blocks identified earlier, we find that *miR159a* and *miR159b* reside within an intrachromosomal duplication within chromosome 1 (Fig. 2A). While many of the conserved genes within the duplicated block appear to maintain their order between the two chromosomal segments, the inversion within the middle of the duplicated region indicates it has undergone an additional rearrangement event. However, the origin of *miR159c* is more mysterious. There are very few conserved genes surrounding *miR159a* and *miR159c* or *miR159b* and *miR159c*, indicating that either *miR159c* arose via a small duplication that did not involve flanking protein-coding genes, that the duplication is ancient and cannot be detected by our methods, or that *miR159c* arose via an unknown mechanism. Overall, we believe that *miR156a* and *miR156b* evolved from a duplication event within the last 30 Myr, while *miR159c* existed prior to this duplication event, as shown in Figure 3A.

The closest downstream MPSS tags for *miR159a* and *miR159b* show slight variations in their tissue expression profiles (Fig. 2D). Under identical conditions, each miRNA demonstrates expression within inflorescence, leaves, root, and silique. However, *miR159a* is expressed in germinating seed, and only *miR159b* is expressed in callus tissue. This example suggests that the duplicated copies exhibit both redundancy of function and diversification. These miRNAs have a wide range of tissue expression and are expressed at low levels; however, there remains the possibility that the MPSS technique failed to detect low levels of expression in callus and seed. Regardless, this does demonstrate the high level of redundant function between the two miRNAs.

miR166 family evolution

Class III HD-ZIP genes are predicted transcription factors that are involved in the adaxial identity of lateral organs and meristem development in *Arabidopsis* (Engstrom et al. 2004; Juarez et al. 2004). The putative binding site, as determined by sequence identity, overlaps with a gain-of-function mutation, suggesting that members of the miRNA166 family regulate these transcription factors (Reinhart et al. 2002; Rhoades et al. 2002). Figure 2B shows two duplicated blocks containing five of the seven miRNA genes within the miR166 family. The first example is of a duplicated block between *miR166a* and *miR166b*, located on chromosomes 2 and 3, respectively. The second example shows the duplicated region between the tandem duplication of *miR166c* and *miR166d* to the chromosomal region surrounding *miR166g*.



miRNA	caf	cas	inf	ins	ap1	ap3	agm	sap	sif	sis	lef	les	so4	s52	rof	ros	gse	TPM
miR166a	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	22
miR166b	0	14	0	0	5	0	0	0	0	0	0	0	0	0	0	0	18	37
miR166d	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
miR166f	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21
miR159a	0	0	0	0	0	0	2	5	32	0	0	0	0	11	4	0	11	65
miR159b	3	7	11	8	5	17	4	11	26	22	9	2	2	1	39	6	0	173
miR156a	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	6
miR156b	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
miR156d	1	0	2	7	0	0	31	29	0	0	43	5	5	0	4	0	0	127
miR156f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	4

Callus | Inflorescence | Silique | Leaves | Root | Germinating Seed

Figure 2. Conserved protein-coding genes surround flanking miRNA genes. The chromosomal regions surrounding two miRNAs are displayed as vertical yellow lines. Each of the protein-coding genes nearby are shown as black horizontal lines, while the miRNA is displayed as a red horizontal line, and indicated by the arrow because of the resolution of the images. The green lines represent genes that are conserved according to BLASTN analysis. (*A*) miR159 family. (*B*) miR166 family. (*C*) miR156 family. (*D*) MPSS tissue expression for miRNA genes from miR166, miR159, and miR156.

Within the highly conserved regions of this intrachromosomal duplication on chromosome 5, some duplicated blocks have undergone smaller inversions and rearrangements.

Differential gene loss after a genome-wide duplication could contribute to a number of miRNA genes that are not visible in the analysis (Paterson et al. 2004). For instance, *miR166c* and *miR166d* are tandemly duplicated, yet there is only one corresponding miRNA, *miR166g*, residing within the duplicated region. The first explanation is that the tandem duplication occurred before the larger duplication event and was followed by differential gene loss near *miR166g*. Alternatively, this could be due to a tandem duplication occurring after a genome-wide duplication event.

To help resolve the evolutionary history of the miR166 family, we looked for conservation in the non-coding flanking regions of the miRNAs. We aligned flanking regions using Dotmatcher from the EMBOSS analysis package (Rice et al. 2000). This demonstrated that there are conserved non-coding regions flanking *miR166b* and *miR166e*, but no regions of conservation between *miR166a* and *miR166e* outside of the conserved stems (Fig. 4A,B,C). The number of conserved regions flanking *miR166b* and *miR166e* is less than the number of regions with sequence similarity between *miR166a* and *miR166b* (Supplemental Fig. 2). This supports the model that the duplication event between *miR166b* and *miR166e* predates the duplication event between *miR166a* and *miR166b*.

The overall evolutionary model we propose for the miR166 family is shown in Figure 3B. *miR166f* lacks any relation, other than having a similar mature miRNA sequence, to all miR166 genes except for *miR166a* with which it has conservation in the opposing stem of the precursor; therefore, we place *miR166f* closest to *miR166a*. We believe *miR166a* and *miR166b* evolved from a recent large-scale duplication event. *miR166b* and *miR166e* have conserved non-coding flanking sequences, while *miR166a* and *miR166e* lack this conservation, indicating that *miR166b* and *miR166e* most likely evolved from a duplication prior to the large-scale duplication event between *miR166a* and *miR166b*. The best explanation is that *miR166e* is anciently related to

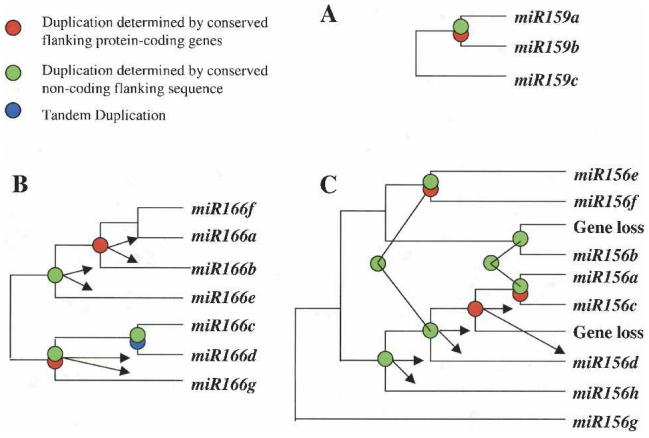


Figure 3. Reconstruction of miRNA family evolution. These phylogenetic trees were generated to demonstrate the order of duplication events for three miRNA families. Circles indicate duplication events for which we have supporting evidence. Red circles indicate duplication events supported by conserved protein-coding genes flanking two miRNA genes. Green circles represent duplication events supported by conserved non-coding sequence flanking two miRNA genes. Blue circles indicate tandemly duplicated miRNAs. A combination of circles indicates that it is supported by multiple methods. The branch lengths are of uniform length and are not meant to indicate time since each duplication event. The connection between two green circles indicates that it is the same duplication event. Arrows establish which two miRNAs were found to be involved in a specific duplication event. (*A*) miR159 family; (*B*) miR166 family; (*C*) miR156 family.

miR166g. *miR166g* resides within a duplicated block with the tandem duplication containing *miR166c* and *miR166d*.

All miR166 family members with an associated MPSS signature demonstrate expression in callus, indicating substantial redundancy of function (Fig. 2D). However, in addition, *miR166a* is expressed in root, and *miR166b* is expressed in germinating seed and inflorescence tissue. *miR166a* and *miR166b* have demonstrated redundant and diversified expression following duplication. Within another duplicated region, *miR166d* (and potentially *miR166c*, depending on whether it resides in the same transcription unit as *miR166d*) has a significant expression level, while its duplicated counterpart, *miR166g*, lacks any detectable level of expression. This either represents the loss of *miR166g* functionality, or indicates that it is transcribed at very low levels indistinguishable from background levels.

The functional implications based on the expression profiles of two tandemly duplicated miRNAs that are located on the same strand is challenging since in many instances only the 3′-miRNA has an associated MPSS signature. For instance, the tandem duplication between *miR166c* and *miR166d*, which resides within the intergenic region between At5g08710 and At5g08720, has one MPSS signature located downstream of the 3′-miRNA, implying that they may be transcribed as one transcriptional unit.

miR156 family evolution

The miR156 family has been demonstrated to target proteins resembling the Squamosa-promoter-binding proteins (SPB). SPB proteins are a plant-specific group of transcription factors involved in plant development (Yamasaki et al. 2004). The complementary target sites for the miRNAs within this family do not reside in the conserved domain defining SPB-like proteins but instead fall within a region weakly conserved among the target family (Bartel 2004).

Figure 2C shows an overview of the relationships between the different members of the family. The different members reside within both inter- and intrachromosomal duplications and appear to occur in pairs (*miR156b*/*miR156c* and *miR156d*/*miR156e*). These closely related pairs are located many genes apart, whereas most pairs that we have characterized as tandem duplications occur within the same intergenic region.

Our overall evolutionary reconstruction (Fig. 3C) shows *miR156g* as an outlier since it has a low level of conservation in the flanking protein-coding genes with *miR156e*, but lacks any other relationship within this family indicating its ancient origins (Supplemental Fig. 2). *miR156h* and *miR156d* have conservation in their flanking non-coding sequence, indicating they have evolved from a duplication event. *miR156b* has conservation across both stems of the precursor with *miR156f*, whereas it only shares similarity in its mature miRNA product with the remainder of the miRNA genes in the family. This suggests an ancient relationship between *miR156b* and *miR156f*. We observed an apparent large-scale duplication involving *miR156e* and *miR156f*. The protein-coding genes conserved in this duplicated block span the region containing *miR156d* (Fig. 2C), yet there isn't a known miRNA in the corresponding region of the duplicated block. We used Patscan to search the region for a miRNA sequence with up to five mismatches that could form a hairpin structure representing a potentially undetected member of the miR156 family but failed to find such a candidate. We therefore believe a gene loss occurred within this region after the duplication event. *miR156d* and *miR156c* were then duplicated from one another based on their conserved flanking protein-coding genes. The most recent duplication event occurred between *miR156a* and *miR156c* as determined by the high level of conserved flanking protein-coding genes. In addition, we think that an ancestor of *miR156b* originally resided within this duplicated block, but once again there were no remnants of a corresponding miRNA within the duplicated block, indicating that the duplication of *miR156b* was again followed by gene loss.

While we lack MPSS expression data for any two miRNAs that are directly involved within a large-scale duplicated region, we do have two miRNAs that are indirectly related according to our evolutionary reconstruction. *miR156c* was involved in a duplication event with *miR156d* prior to its recent duplication with *miR156a*. Interestingly, we do not have an MPSS signature for *miR156c*, but we do have a signature for *miR156a* and *miR156d* (Fig. 3D). We observed a broad expression profile (callus, inflorescence, leaves, and root) for the more divergent *miR156d* and a very specific expression profile (leaves) for *miR156a*. This suggests that *miR156a* is providing redundant functionality with *miR156d*, while *miR156c* may have lost some functionality following its duplication with *miR156a*.

miR395 family evolution

The miR395 family, predicted to target mRNAs coding for ATP sulphurylases, can be broken into two groups of tandem duplications (Bartel 2004; Jones-Rhoades and Bartel 2004). Each group of tandemly duplicated miRNAs has two miRNAs on the same strand and another on the opposite strand. Our previous expansion analysis indicated that seven protein-coding genes flanking both sets of tandemly duplicated miR395 genes were conserved. This suggests that an intrachromosomal duplication event occurred after the tandem duplication events, thereby conserving the orientation of the miR395 genes (Fig. 5).

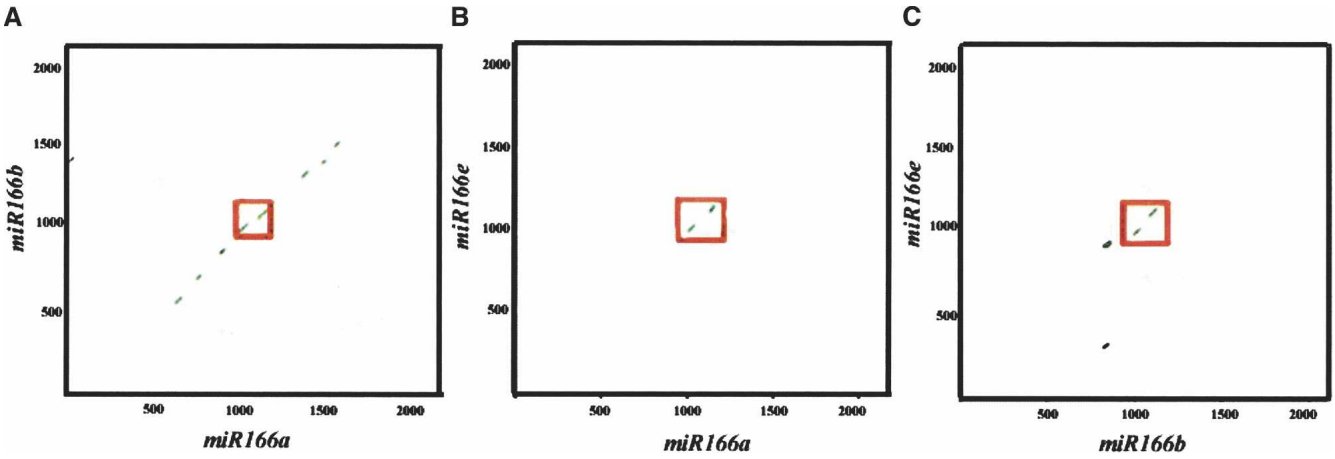


Figure 4. Dotmatcher results for miR166 family. These three plots highlight non-coding flanking regions that are conserved between miRNAs. The red boxes highlight the conserved stems between the two miRNAs. (*A*) *miR166a* and *miR166b*. (*B*) *miR166a* and *miR166e*. (*C*) *miR166b* and *miR166e*.

Within both sets of tandem duplications, we observed a high sequence complementarity within the loop regions, providing further support that each set arose from tandem duplication events. One example of two highly conserved precursors within the miR395 family is between *miR395b* and *miR395c*. Both miRNAs are in the same orientation and have an identical precursor length of 100 nt, yet only two nucleotides within their loop regions are different.

In the other set of tandem duplications among *miR395d*, *miR395e*, and *miR395f*, the two miRNAs on the same strand also have a higher level of similarity in their loop region than they do with the miRNA on the opposing strand. Regardless, they all have a high level of sequence conservation, being tandem duplications of one another.

Only *miR395e* has an associated MPSS signature, making it difficult to draw any conclusions about potential diversification. The expression of miR395, which depends on environmental stress, increases during sulfate starvation (Bartel 2004; Jones-Rhoades and Bartel 2004). The specificity of this condition makes miR395 less likely to appear in tissue libraries tested with MPSS. According to the MPSS data, in the instance of *miR395e*, the tandem duplications do not appear to be transcribed as one transcriptional unit, given that the signature is located downstream of *miR395e*, which was not the 3′ member of the pair.

Discussion

The evolution of protein-coding genes arises from genome-wide duplication events, large-scale chromosomal duplication, and local rearrangements. Recent efforts in miRNA predictions provide a solid foundation for analyzing the evolution of miRNAs. By analyzing the genomic position of known miRNA families, we demonstrate that miRNAs evolve through segmental duplications and tandem duplications in the same manner as protein-coding genes.

Five of the six sets of tandemly duplicated miRNAs that we observed are in arrays or two or three miRNAs, which is in agreement with the observation that 87% of all tandemly duplicated *Arabidopsis* protein-coding genes occur in arrays or two or three genes (Zhang and Gaut 2003) due to shrinkage of the genome over the last 50 Myr. In addition, ~17% of all *Arabidopsis* protein-coding genes reside within tandemly repeated segments (Vision et al. 2000), which is slightly lower than that of miRNAs at 25%.

For large-scale duplications, we observed a higher rate of intrafamily duplicated blocks than we did for randomly selected locations or for miRNAs from different families. In addition to seeing a higher rate of apparent duplicated blocks surrounding miRNAs from the same families, the level of conservation of the flanking proteins was generally higher within miRNA families than duplicated blocks surrounding randomly selected locations and miRNAs from different families. Two of the duplications having at least four or more conserved flanking protein-coding genes (*miR159a*/*miR159b* and *miR166a*/*miR166b*) were also found in the initial study of large-scale duplications conducted by the *Arabidopsis* Genome Initiative (2000). This demonstrates that duplication events have caused miRNA family expansion just as they have for protein-coding genes.

A total of 59 (67%) multifamily miRNA genes were within either a tandem or large-scale duplication. We believe that miRNAs not occurring within duplicated regions are the result of older, less detectable, duplication events, rather than random insertions. The accumulation of chromosomal rearrangements over time, in addition to events such as gene loss, are some of the more well-known hindrances to detecting older duplications, and therefore may limit our findings to more recently duplicated miRNAs (~39 Myr).

Our understanding of miRNA evolution serves as a starting point for elucidating their complex regulatory roles. Expression data provide some insight into the functional divergence of duplicated miRNAs by capturing differences in specific tissue samples. We chose to use the MPSS expression data set because it can distinguish between different miRNA loci and has 17 different tissue-specific libraries for comparing expression profiles. Additional large-scale expression data sets such as ESTs or cloned libraries were too limiting to incorporate into our analysis. Only two miRNAs were captured via ESTs. The ASRP data set (Gustaf-

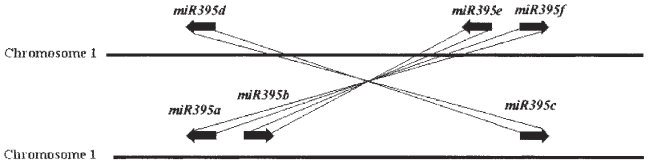


Figure 5. Schematic representation of intrachromosomal duplication within the miR395 family.

son et al. 2005) is highly sensitive, but fails to distinguish among family members.

The cutoff that we used to determine whether a downstream MPSS signature should be associated with a miRNA is arbitrary. Supporting our cutoff choice, we were able to observe the characteristically higher density of signatures slightly downstream, ~400 nt, of the miRNA precursor. This observation is consistent with previous work done on public and private MPSS sets in which the majority of miRNAs had a signature within 500 nt downstream of the miRNA (Wang et al. 2004). To provide further evidence that these downstream signatures are in fact representing the miRNA transcript, we looked at ESTs. Due to the lack of *Arabidopsis* ESTs containing miRNAs, available ESTs from other plant species provide evidence for expression downstream of the precursor (Bonnet et al. 2004; Xie et al. 2005; Zhang et al. 2005). Therefore, the downstream MPSS signatures are associated with the miRNA transcript.

Data from MPSS detect just over a third of all known miRNAs. While this number may appear low, it still provides locus-specific expression data. Many of the miRNA loci that were not captured by MPSS were also missed by a combination, if not all, of experimental methods such as cloning, 5′-RACE, and 3′-RACE. This indicates that many of these miRNA genes have low or very cell-specific expression.

Using expression data beyond validating miRNA existence is a challenging task and is limited by the sampling of the tissues at specific points in time. While presence of miRNA expression is informative, absence of expression must be interpreted cautiously. The tissues that lack expression may result from low expression levels, sensitivity, or limitations of the assay. Therefore, the expression data serve as a good starting point for understanding the expression patterns within miRNA families, but will need to be expanded on to have a true understanding of the temporal and spatial patterns of miRNA genes.

Within animal species, miRNAs are commonly found in clusters in which multiple miRNAs are transcribed at the same time in one large polycistronic unit. Consistent with this is our observation that for three tandem duplications in the *Arabidopsis* genome in which the miRNA is found in the same orientation, there is a single associated MPSS signature downstream of the 3′-miRNA. In these instances, the 5′-miRNA lacked a signature with a significant level of expression. An alternative explanation is simply the lack of expression of the 5′-miRNA.

Overall, we have demonstrated that plant miRNAs families are evolving through duplication events similar to those that drive the evolution of protein-coding genes, and that the duplicated copies take on new expression patterns potentially resulting in neo- and subfunctionalization. The evolutionary relationships within a miRNA family in conjunction with public data enable us to explore the subsequent functional divergence of duplicated genes and can be used for further experimental analysis of their interactions with target mRNA and resulting regulatory effects in plant development. While we have documented specific examples of divergent expression profiles following a duplication event, a more comprehensive understanding will become clear as more expression data become available within *Arabidopsis*. Our procedures can also be applied in other cereal species, which contain similar families to *Arabidopsis*, and some monocot-specific families. On a more practical note, our understanding and ability to control gene expression during plant development have the potential to improve crop yields, increase resistance to disease, and increase the adaptability of the plant to its environment. The ability to understand the evolution of plant miRNAs will enable us to understand the complexities of miRNA-based regulation.

Methods

Identification of miRNA genes

To determine the genomic locations of miRNA genes, we downloaded miRNA sequences from the miRNA Registry version 5.0 (http://www.sanger.ac.uk/cgi-bin/Rfam/mirna/), a database of published miRNAs (Griffiths-Jones 2004), and aligned them against the TIGR *Arabidopsis* genome version 5.0. The protein-coding genes flanking each miRNA were then extracted for our miRNA family expansion analysis.

Categorization of miRNA expansions

For this analysis, we focused on the processes of segmental and tandem duplication, using the similarity among sets of protein-coding genes as markers for regions involved in such duplications (Vision et al. 2000). To categorize apparent expansions of miRNA gene families, we looked at the physical locations of all the members of a family. Tandem duplications are characterized as multiple members occurring within the same intergenic region, or within neighboring intergenic regions.

In order to classify two miRNAs as residing within a duplicated block, their neighboring protein-coding genes must have high similarity to one another at the amino acid level. Therefore, to identify segmental duplications, we developed Perl scripts that extract 10 protein-coding genes upstream and downstream of each miRNA, or tandemly duplicated miRNAs since their flanking protein-coding genes would be the same. The protein-coding genes flanking each miRNA were aligned against a set of 29,161 *Arabidopsis* peptide sequences (http://www.arabidopsis.org) at the amino acid level, using BLASTP, to retrieve the best non-self matches (assuming *E*-value < 0.001). For each miRNA, we tallied the number of flanking protein-coding genes with a best non-self match to a protein-coding gene neighboring a miRNA from the same family (i.e., *miR156a* and *miR156b*).

Simulation of miRNA expansions

We generated a simulation to determine the random likelihood of a protein-coding gene flanking a miRNA to have a best match to a protein-coding gene neighboring a related miRNA. The simulation randomly selected two protein-coding genes as anchors, representing two related miRNAs from the same family, and then aligned the 10 flanking protein-coding genes against all *Arabidopsis* genes using BLASTP. It then tallied the total number of protein-coding genes from the first anchor that had a best non-self match (*E*-value < 0.001) with a protein-coding gene neighboring the other anchor point. We repeated this process 1000 times to recreate the frequency of observing a duplication event between two genomic regions.

Estimation of synonymous substitutions and duplication event dating

A Perl script parsed the peptide alignment, from BLASTP, for each pair of conserved flanking protein-coding genes within each miRNA duplicated region to obtain a high quality alignment. Using the protein alignment as our guide, the codons were extracted for each amino acid that was aligned between genes, excluding regions containing gaps. The level of synonymous substitution for these nucleotide sequences was calculated with codeml (Yang 1997), which uses a maximum likelihood method

under the F3×4 model (Goldman and Yang 1994). The mean K_s value was calculated for each pair of protein-coding genes within a duplicated block and then used for determining the approximate date of divergence, D, with the equation: $D = K_s/2E$. We assumed a constant rate of synonymous substitution for dicots, E, as 1.5×10^{-8} substitutions/synonymous site/year (Koch et al. 2000).

Expression analysis

We obtained MPSS signatures from the Delaware Biotechnology Institute (http://mpss.udel.edu/at/). All of the MPSS signatures were loaded into a custom MySQL database designed for this task. The intergenic region downstream of each miRNA was extracted and then scanned for *dpn-II* restriction sites, used by the MPSS technology. For each *dpn-II* site, the 20-mer signature was extracted and queried against our database to filter out all signatures lacking reliability, uniqueness, or a significant expression level. Each signature is grouped into a class indicating the signature position relative to the genome annotation. Only class 4 signatures, indicating transcript expression within an intergenic region, were extracted. We associated the first downstream signature meeting these criteria with the miRNA.

Acknowledgments

We thank T. Kellog and N. Chen for their critical reading of the manuscript and K. Nabuta and B. Meyers for the MPSS data. This work was supported by the National Science Foundation (grants #0321685 and #27870201) and USDA ARS CRIS project 1907-21000-014.

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.

Adams, K.L. and Wendel, J.F. 2005. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8:** 135–141.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., et al. 2003. A uniform system for microRNA annotation. *RNA* **9:** 277–279.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

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.

Blanc, G. and Wolfe, K.H. 2004a. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16:** 1679–1691.

———. 2004b. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16:** 1667–1678.

Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. 2004. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc. Natl. Acad. Sci.* **101:** 11511–11516.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422:** 433–438.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D., and May, G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* **2004:** 4–10.

Emery, J.F., Floyd, S.K., Alvarez, J., Eshed, Y., Hawker, N.P., Izhaki, A., Baum, S.F., and Bowman, J.L. 2003. Radial patterning of *Arabidopsis* shoots by class III HD-ZIP and KANADI genes. *Curr. Biol.* **13:** 1768–1774.

Engstrom, E.M., Izhaki, A., and Bowman, J.L. 2004. Promoter bashing, microRNAs, and Knox genes. New insights, regulators, and targets-of-regulation in the establishment of lateral organ polarity in *Arabidopsis*. *Plant Physiol.* **135:** 685–694.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11:** 725–736.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res.* **32:** D109–D111.

Gustafson, A.M., Allen, E., Givan, S., Smith, D., Carrington, J.C., and Kasschau, K.D. 2005. ASRP: The *Arabidopsis* Small RNA Project Database. *Nucleic Acids Res.* **33:** D637–D640.

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.

Juarez, M.T., Kui, J.S., Thomas, J., Heller, B.A., and Timmermans, M.C. 2004. microRNA-mediated repression of rolled leaf1 specifies maize leaf polarity. *Nature* **428:** 84–88.

Koch, M.A., Haubold, B., and 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.

Lai, E.C., Tomancak, P., Williams, R.W., and Rubin, G.M. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4:** R42.

Lawton-Rauh, A. 2003. Evolutionary dynamics of duplicated genes in plants. *Mol. Phylogenet. Evol.* **29:** 396–409.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.* **23:** 4051–4060.

Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14:** 1605–1619.

Maher, C., Timmermans, M., Stein, L., and Ware, D. 2004. Identifying microRNAs in plant genomes. In *Computational systems bioinformatics* (ed. IEEE), (ed. F. Titsworth), pp. 718–723. IEEE, Stanford, CA.

Meyers, B.C., Tej, S.S., Vu, T.H., Haudenschild, C.D., Agrawal, V., Edberg, S.B., Ghazal, H., and Decola, S. 2004a. The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. *Genome Res.* **14:** 1641–1653.

Meyers, B.C., Vu, T.H., Tej, S.S., Ghazal, H., Matvienko, M., Agrawal, V., Ning, J., and Haudenschild, C.D. 2004b. Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. *Nat. Biotechnol.* **22:** 1006–1011.

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.

Papp, I., Mette, M.F., Aufsatz, W., Daxinger, L., Schauer, S.E., Ray, A., van der Winden, J., Matzke, M., and Matzke, A.J. 2003. Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant Physiol.* **132:** 1382–1390.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci.* **101:** 9903–9908.

Prince, V.E. and Pickett, F.B. 2002. Splitting pairs: The diverging fates of duplicated genes. *Nat. Rev. Genet.* **3:** 827–837.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. 2002. MicroRNAs in plants. *Genes & Dev.* **16:** 1616–1626.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. 2002. Prediction of plant microRNA targets. *Cell* **110:** 513–520.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16:** 276–277.

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

Vision, T.J., Brown, D.G., and Tanksley, S.D. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* **290:** 2114–2117.

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.

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

Yamasaki, K., Kigawa, T., Inoue, M., Tateno, M., Yamasaki, T., Yabuki, T., Aoki, M., Seki, E., Matsuda, T., Nunokawa, E., et al. 2004. A novel zinc-binding motif revealed by solution structures of DNA-binding domains of *Arabidopsis* SBP-family transcription factors. *J. Mol. Biol.* **337:** 49–63.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.

Zhang, L. and Gaut, B.S. 2003. Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? *Genome Res.* **13:** 2533–2540.

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.

Received September 22, 2005; accepted in revised form December 29, 2005.