GBE

Genome Biology and Evolution

SMBE

Origins and Evolution of MicroRNA Genes in Plant Species

Masafumi Nozawa[1,*], Sayaka Miura, and Masatoshi Nei

Department of Biology, Institute of Molecular Evolutionary Genetics, Pennsylvania State University

[1]Present address: Division of Evolutionary Biology, National Institute for Basic Biology, 38 Nishigonaka, Myodaiji, Okazaki, Aichi 444-8585, Japan.

*Corresponding author: E-mail: nozabey@nibb.ac.jp.

Accepted: 20 December 2011

Abstract

MicroRNAs (miRNAs) are among the most important regulatory elements of gene expression in animals and plants. However, their origin and evolutionary dynamics have not been studied systematically. In this paper, we identified putative miRNA genes in 11 plant species using the bioinformatic technique and examined their evolutionary changes. Our homology search indicated that no miRNA gene is currently shared between green algae and land plants. The number of miRNA genes has increased substantially in the land plant lineage, but after the divergence of eudicots and monocots, the number has changed in a lineage-specific manner. We found that miRNA genes have originated mainly by duplication of preexisting miRNA genes or protein-coding genes. Transposable elements also seem to have contributed to the generation of species-specific miRNA genes. The relative importance of these mechanisms in plants is quite different from that in *Drosophila* species, where the formation of hairpin structures in the genomes seems to be a major source of miRNA genes. This difference in the origin of miRNA genes between plants and *Drosophila* may be explained by the difference in the binding to target mRNAs between plants and animals. We also found that young miRNA genes are less conserved than old genes in plants as well as in *Drosophila* species. Yet, nearly half of the gene families in the ancestor of flowering plants have been lost in at least one species examined. This indicates that the repertoires of miRNA genes have changed more dynamically than previously thought during plant evolution.

Key words: birth-and-death evolution, gene duplication, multigene family, small RNA, transposable element.

Introduction

MicroRNAs (miRNAs) regulate gene expression of a variety of protein-coding genes at the posttranscriptional level (Bartel 2004, 2009; Voinnet 2009). They are first transcribed as primary miRNAs with a few hundred nucleotides, but after several processing steps, the mature miRNA sequences consisting of only 21–24 nt are formed (Shabalina and Koonin 2008; Bartel 2009; Carthew and Sontheimer 2009). Whereas the seed sequence (second to seventh nucleotides of the mature sequence) is most important for recognizing the transcript of target protein–coding genes in animals, the entire mature sequence is used for recognizing the transcript with near-perfect base pairings in plants (Axtell and Bowman 2008).

Most plant species contain more than 100 miRNA genes in their genome (e.g., Axtell et al. 2007; Fahlgren et al. 2007; Zhu et al. 2008; Klevebring et al. 2009; Zhang et al. 2009; Joshi et al. 2010; Ma et al. 2010; Dhandapani et al. 2011). As to the evolutionary origin of miRNA genes, there are several different mechanisms proposed. First, miRNA genes may be generated from duplicates of protein-coding genes. This is an attractive hypothesis because an miRNA gene generated from a protein-coding gene would bind to the transcript of the protein-coding gene. In fact, there seem to be a number of miRNA genes generated in this way (Allen et al. 2004; Rajagopalan et al. 2006; Fahlgren et al. 2007, 2010). Second, transposable elements (TEs) may become miRNA genes. In particular, miniature inverted-repeat transposable elements (MITEs) have a potential to become miRNA genes because they have inverted repeats with a short internal sequence, which can potentially turn into the hairpin structure of an miRNA gene. It has been proposed that dozens of miRNA genes originated from MITEs or other TEs in *Arabidopsis* and rice (Piriyapongsa and Jordan 2008). Third, new miRNA genes may be generated by duplication of preexisting miRNA genes with subsequent mutations. This mechanism seems to be important in plants because each miRNA gene family

on average consists of several miRNA genes (Li and Mao 2007). The last mechanism is that miRNA genes naturally arise by spontaneous mutations from hairpin structures in the genome. It has been hypothesized that some miRNA genes have been generated in this way in *Arabidopsis* (De Felippes et al. 2008).

The relative roles of these mechanisms in plants have been studied by the comparative genomics approach. Using the two closely related species, *Arabidopsis thaliana* and *A. lyrata*, which diverged ~10 Ma, Fahlgren et al. (2010) proposed that a large proportion of young miRNA genes originated by duplication of protein-coding genes. However, they also indicated that most of these miRNA genes are evolutionarily short-lived and unlikely to become functionally relevant (see also Ma et al. 2010). Cuperus et al. (2011) also reviewed the recent progress about the evolution of miRNA genes in plants. However, the relative importance of these mechanisms for the origin of evolutionarily long-lived miRNA genes remains unclear. Therefore, we need to analyze the miRNA genes in more details by comparing distantly related species to obtain a general idea about the origins of miRNA genes in plants.

In our previous study (Nozawa et al. 2010), we investigated the origins and evolutionary dynamics of miRNA genes in 12 *Drosophila* species by using the bioinformatic technique and suggested that many miRNA genes have originated from hairpin structures in introns or intergenic regions (see also Lu et al. 2008), although duplication of preexisting genes has apparently been important as well. We also found that the old miRNA genes that originated before the divergence of *Drosophila* species (>60 Ma) have been under strong functional constraints, whereas young miRNA genes generated ~10 Ma evolve in a more or less neutral fashion. Because extensive data on genomic and miRNA gene sequences are now available from flowering plants, moss, and green algae, we have applied the same technique to study the evolutionary changes of miRNA genes in plants and compared them with those in *Drosophila* species.

Materials and Methods

Identification of miRNA Genes in Plant Species

The genome sequences from *Arabidopsis* (*A. thaliana*), papaya (*Carica papaya*), poplar (*Populus truncatula*), soybean (*Glycine max*), grape (*Vitis vinifera*), rice (*Oryza sativa*), *Sorghum* (*S. bicolor*), maize (*Zea mays*), moss (*Physcomitrella patens*), and green algae (*Chlamydomonas reinhardtii*) were downloaded from Phytozome version 6.0 (http://www.phytozome.net). The *Medicago* (*M. truncatula*) genome sequence (Mt3.0) was downloaded from Medicago.org (http://medicago.org).

For obtaining the query sequences for homology search, we downloaded all miRNA genes from the miRBase (Release 16, http://www.mirbase.org, Griffiths-Jones et al. 2008) in *Arabidopsis*, *Medicago*, grape, rice, maize, moss, and green algae, where small RNA sequencing data are available (for Gene Expression Omnibus accession nos., see supplementary table S1, Supplementary Material online). We then applied the standard criteria proposed by Meyers et al. (2008) to these sequences to eliminate spurious miRNA genes as much as possible. In short, the criteria were 1) typical secondary structures of miRNAs as defined in the software MIRcheck (Jones-Rhoades and Bartel 2004), 2) $\geq$10 small RNA reads of the mature sequence or $\geq$1 reads of both mature and star sequences, 3) $\geq$0.25 for the proportion of the number of mature and star reads out of the total number of reads mapped anywhere on the hairpin sequence, and 4) $\geq$3 for the ratio of the number of reads mapped anywhere on the hairpin to the number of reads mapped anywhere on the complementary strand of the hairpin. Here, the star sequence means the sequence that forms a duplex with the mature sequence. In this way, we obtained 758 miRNA genes that satisfy all the criteria (supplementary table S2, Supplementary Material online). These genes were classified into gene families based on the information given in the miRBase. The mature sequences of miRNA genes within families showed $\geq$85% sequence identity in most cases (i.e., the number of nucleotide differences between the sequences was $\leq$3), whereas the sequences between different gene families were mostly unalignable.

Using these 758 miRNA hairpin sequences as queries, we conducted a BLASTN search against each genome sequence with the cutoff *E* value of 10^{-3}. We then extracted hit sequences with 300 nt of the 5′ and 3′ flanking sequences from a genome. If an extracted sequence contained a mature-like sequence ($\leq$2 nt mismatches with the mature sequence of the query miRNA without any indels) and had a typical secondary structure of miRNAs defined in the MIRcheck, the sequence was regarded as a putative miRNA gene. Finally, to classify miRNA genes into TE-like and non-TE-like miRNA genes, we used the RepeatMasker (open-3.2.8). The RepBase15.10 (http://www.girinst.org) and the Plant Repeat Database at Michigan State University (http://plantrepeats.plantbiology.msu.edu) were used for the databases of the RepeatMasker. The nucleotide sequences and genomic locations of the miRNA genes are given in the miRNA_hairpin.fas (Supplementary Material online) (or miRNA_mature.fas, Supplementary Material online for mature sequences) and supplementary table S3 (Supplementary Material online), respectively.

Estimation of the Numbers of miRNA Genes in Ancestral Species

To estimate the numbers of miRNA genes in the ancestral species and the numbers of gene gains and losses in the evolutionary process, we used the following parsimony method. Suppose there is an miRNA gene family which exists in *Arabidopsis* (2 genes), *Medicago* (1 gene), soybean (1 gene), grape (1 gene), rice (1 gene), *Sorghum* (1 gene),

and maize (1 gene) among nine species of flowering plants examined but not in moss and green algae (supplementary fig. S1, Supplementary Material online). Because the phylogenetic tree for the 11 species examined here is known, we can easily infer the origin and evolutionary changes of the genes belonging to this family (supplementary fig. S1, Supplementary Material online). That is, this gene family originated in the ancestral lineage of flowering plants after the lineage diverged from the moss lineage. However, the gene was later duplicated into two copies in the *Arabidopsis* lineage and was lost in the papaya and poplar lineages. If we construct the evolutionary changes of miRNA genes for all gene families, we can estimate the total numbers of miRNA genes in all the ancestral species and their evolutionary changes. Here, we did not use the reconciled-tree method (Goodman et al. 1979; Nam and Nei 2005; Niimura and Nei 2007) because the nucleotide sequences of miRNA genes were too short to make reliable gene trees.

Estimation of Substitution Rates of miRNA Genes

To study the rate of nucleotide substitution in the mature region of miRNA genes, we aligned mature sequences for each gene family by using MUSCLE (Edgar 2004). In this analysis, we used only the miRNA sequences in *Arabidopsis*, papaya, poplar, *Medicago*, and soybean. In other words, only the divergence time of 109 Ma between the *Arabidopsis*–papaya and the poplar–*Medicago*–soybean (Hedges and Kumar 2009) was used for the estimation. This approach made it possible to remove the potential bias in comparing the substitution rates of different gene families, which were generated in different times of plant evolution. When a species has two or more homologous genes for a given gene family, we used all homologous genes and computed the average sequence divergence between species. We first computed the average Jukes and Cantor (1969) distance ($\bar{D}$) between the sequences from *Arabidopsis* (or papaya) and poplar (*Medicago*, or soybean) for each miRNA gene family (i.e., the sequence divergence between groups I and II in supplementary fig. S1, Supplementary Material online). The substitution rate (R) was then computed by $R=\bar{D}/2T$, where T is the divergence time (109 Myr in this case). In this way, we computed the substitution rates for 24 miRNA gene families. We finally computed the average substitution rate for these gene families for each tree branch. Furthermore, using the hairpin structures predicted by the software RNAFOLD (Mathews et al. 1999), we extracted the star sequences and computed the substitution rates for them.

We also computed the rates of synonymous (r_S) and nonsynonymous (r_N) substitutions for protein-coding genes. For this analysis, all annotated coding sequences (CDS) for each species were downloaded from the same databases mentioned above. We then conducted reciprocal BLASTP searches between all pairs of sequences for different species with a cutoff E value of 10^{-5} and determined the orthologous relationships of all genes. For the computation of r_S and r_N, we used the modified Nei–Gojobori method (Zhang et al. 1998) with a transition/transversion ratio of 1.5 (Morton et al. 2006). (Perl scripts are available upon the request to M. Nozawa.) We used the 7,147 orthologous genes with ≥100 alignable codons for this analysis.

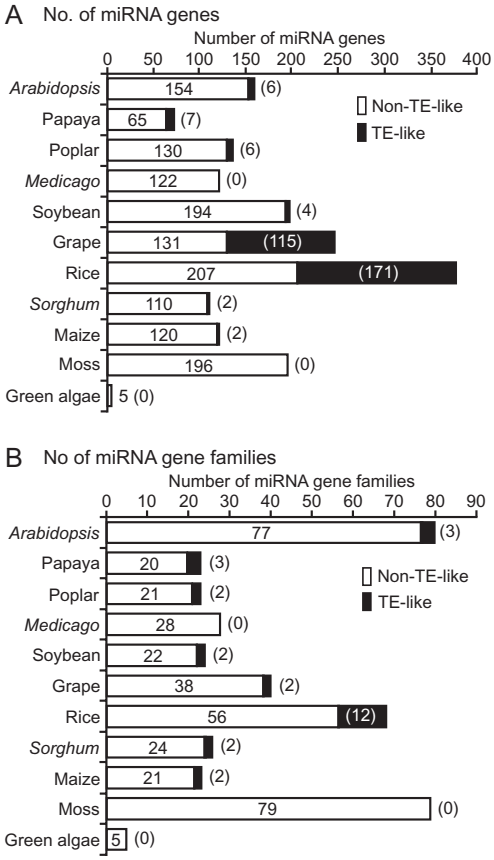

Fig. 1.—Numbers of (*A*) miRNA genes and (*B*) gene families in the 11 plant species. White and black bars represent the numbers of non-TE-like and TE-like (in parentheses) miRNA genes (or gene families), respectively.

Results

Numbers of miRNA Genes

Our homology search identified ~1,400 non-TE-like and ~300 TE-like miRNA genes in the 11 plant species (fig. 1*A*). The number of miRNA genes in green algae was only five, although 50 "potential" miRNA genes have been deposited in the miRBase. This difference occurred because we used stringent criteria for identifying miRNA genes to avoid false identification. It should be noted, however, that even if all 50 genes were used in our analysis, the results were essentially the same as will be shown below. In other species, the number of miRNA genes varied considerably

with species. For example, the number was 72 in papaya, whereas it was as many as 378 in rice. The number of non-TE-like miRNA genes was much greater than the number of TE-like miRNA genes in most species except in grape and rice, where the numbers were nearly the same.

We also counted the number of miRNA gene families in each species. In this study, each gene family was defined as a group of miRNA genes with at least ~85% mature sequence identity based on the classification in the miRBase. The results showed that the number of miRNA gene families (fig. 1*B*) is much smaller than the number of miRNA genes (fig. 1*A*), as was noted previously (Li and Mao 2007). This suggests the importance of gene duplication for generating new miRNA genes. Note that the number of TE-like miRNA gene families is much smaller than the number of non-TE-like miRNA gene families even in grape and rice, suggesting that only a small number of gene families consist of TE-like miRNA genes in these species.

It should be noted that when all plant miRNA genes deposited in the miRBase were used for the homology search without any filtration, the number of potential miRNA genes identified increased especially for TE-like miRNA genes (data not shown). However, the number of miRNA gene families did not increase appreciably, and the general evolutionary pattern remained unchanged. Therefore, we believe that our results are robust irrespective of the definition of miRNA genes.

Chromosomal Distributions

In all the species examined, miRNA genes are scattered throughout the genome (supplementary table S3, Supplementary Material online). A majority (84%) of miRNA genes are located in intergenic regions (supplementary table S4, Supplementary Material online). This situation is contrary to that of *Drosophila* species, in which nearly half of the miRNA genes are located in introns (Nozawa et al. 2010). It should be noted that the proportion of intergenic miRNA genes is not correlated with the genome size of the plant species examined ($R = 0.27$, $P = 0.42$, supplementary fig. S2, Supplementary Material online). In fact, *Arabidopsis* shows a high proportion of intergenic miRNA genes (86%) even though its genome size is as small as that of *D. melanogaster*. We also found that 76 (4%) miRNA genes are overlapping with CDS on the chromosomes of the 11 plant species, but many of them could be annotation errors. Indeed, only two (1%) miRNA genes are located within CDS in the *Arabidopsis* genome, which is by far the best annotated in plants.

To examine the homologous relationships of miRNA genes between *Arabidopsis* and rice, we determined their chromosomal locations (fig. 2). Our analysis revealed that nearly half of miRNA genes in *Arabidopsis* have their homologous miRNA genes in rice and vice versa (short black rods in fig. 2). Each of these genes in one species is generally related to several genes in the other, indicating that there are many paralogous genes in each species. Because these paralogous genes are frequently located on different chromosomes in the two species, there must have been extensive chromosomal rearrangements after duplication of miRNA genes. The remaining miRNA genes do not have any homologous genes in another species (red rods in fig. 2), suggesting that there are many lineage- or species-specific miRNA genes. This extensive changes of miRNA gene repertoire may be reasonable if we consider that these two species diverged ~150 Ma (Hedges and Kumar 2009).

To examine the importance of tandem duplication for generating new miRNA genes in the genome, we also computed the proportion of genes clustered in each genomic region. Here, a clustered gene is the miRNA gene that is located within 5 kb from at least one other miRNA gene on the chromosomes. The results showed that the proportion of clustered genes is on average only 17% (supplementary table S5, Supplementary Material online), which is much smaller than the proportion (~40%) in *Drosophila* species (Nozawa et al. 2010). We originally thought that because the genome sequences of some plant species such as papaya consist of fragmented scaffolds due to incomplete sequencing, the proportion of clustered genes may be underestimated. However, there was no significant correlation between the number of scaffolds (or chromosomes) and the proportion of clustered genes ($R = 0.11$, $P = 0.74$). This suggests that the results obtained are not artifacts, but miRNA genes are indeed scattered more in plant genomes compared with *Drosophila* genomes. Yet, this does not necessarily mean that tandem duplication is unimportant in the evolution of plant miRNA genes because many genes within a cluster belong to the same gene family. In fact, our study showed that about 92% of gene gains within a cluster belong to the same gene family in plants (supplementary table S5, Supplementary Material online) in contrast to ~30% in *Drosophila* species (Nozawa et al. 2010). Therefore, it appears that many miRNA genes in plants have been generated by tandem duplication of preexisting miRNA genes and then scattered throughout the genome by chromosomal rearrangements.

miRNA Genes Derived from TEs

We next examined what kinds of TEs have generated potentially functional miRNA genes in plants by using the software RepeatMasker. The results showed that miRNA genes from 19 different miRNA gene families have sequence similarity to TEs (supplementary table S6, Supplementary Material online). Therefore, ~8% (19/226) of miRNA gene families may have been derived from TEs. For example, a DNA transposon, *Stowaway*, appears to have contributed to the generation of *MIR-812*, *MIR-818*, *MIR-1862*, and *MIR-1867* gene families, all of which are rice-specific miRNA gene families. This is reasonable because *Stowaway* is one of the most

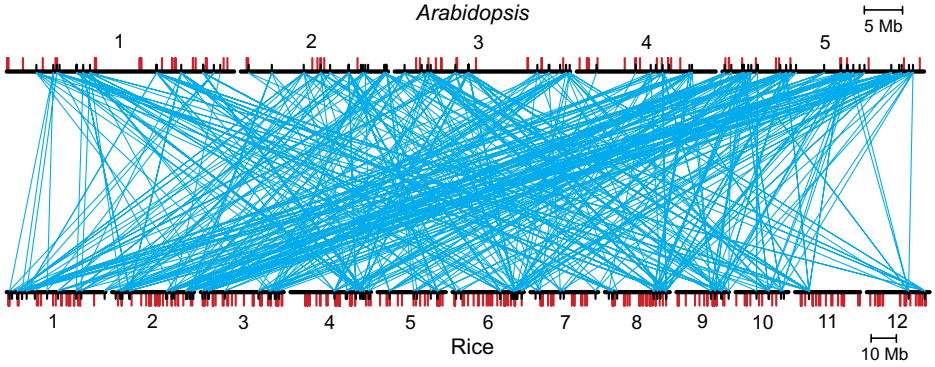

Fig. 2.—Chromosomal locations of miRNA genes and their homologous relationships in *Arabidopsis* and rice. Short black rods on the chromosomes represent the miRNA genes that have at least one homolog in another species, whereas long red rods indicate the miRNA genes that have no homolog in another species. Blue lines illustrate homologous relationships of miRNA genes between *Arabidopsis* and rice. All miRNA genes belonging to the same gene families are connected.

abundant MITEs in the rice genome (Oki et al. 2008). Because 12 of the 19 TE-like miRNA gene families are species-specific, TEs are unlikely to be a major source for the origin old miRNA gene families.

To see whether these TE-like miRNA genes are really functional, we surveyed small RNA data sets (supplementary table S1, Supplementary Material online) and examined the expression of these miRNA genes. We found that 21 different mature miRNAs that are produced from TE-like miRNA genes are highly expressed (≥10 reads). Therefore, it is quite possible that these miRNA genes are indeed functional. For other TE-like miRNA genes, we could not find solid evidence for gene expression. In practice, however, it is quite difficult to distinguish miRNA genes from real TEs because many TEs with inverted repeats can potentially form hairpin structures like miRNA genes (supplementary fig. S3, Supplementary Material online). We therefore excluded all TE-like miRNA genes in the following analyses unless otherwise stated.

Birth-and-Death Evolution of miRNA Genes

We next examined how non-TE-like miRNA genes have evolved during plant evolution. First, to obtain a detailed picture of evolutionary dynamics of non-TE-like miRNA genes, we estimated the minimum numbers of miRNA genes in ancestral species and the numbers of gene gains and losses during plant evolution by using the parsimony method (for details, see Materials and Methods).

Our homology search identified no homologous miRNA genes between green algae and land plants (fig. 3*A*) as previously implied (see Axtell and Bartel 2005; Axtell 2008; Fahlgren et al. 2010). This is true even when all 50 green algae genes deposited in the miRBase were used for the homology search. However, this does not necessarily mean that no miRNA genes existed before the divergence of green algae and land plants because the ancestral miRNA genes may have been lost during plant evolution. After the divergence from green algae, the number of miRNA genes has increased in the land plant lineages, and the number was estimated to be 82 in the ancestor of flowering plants. The number of gene families also increased during this period from 0 to 25, but the extent of the increase was less than that of gene number (fig. 3*B*). This suggests that many gene gains during this period were caused by duplication of miRNA genes. After this period, however, the evolutionary change of the number of genes has varied considerably among the evolutionary lineages. For example, rice has gained 126 genes and lost 1 gene after the divergence of flowering plants, whereas the papaya gained only 8 genes and lost as many as 25 genes during the same period (fig. 3*A*). The same trend was observed for gene families (fig. 3*B*). Note that the numbers of miRNA genes in the ancestral species may be underestimated because some miRNA genes in the ancestors may be extinct from all extant species. Note also that many miRNA genes were estimated to have been gained in the exterior branches leading to the extant species. This occurred partly because we used the parsimony method to estimate the minimum numbers of genes in ancestral species. However, the number of gene families did not always increase in the exterior branches except for *Arabidopsis*, rice, and moss. These results suggest that the number of genes has increased mainly by gene duplication in recent years. This trend was essentially the same even when TE-like miRNA genes were included in the analysis (supplementary fig. S4, Supplementary Material online).

To understand the importance of gene duplication more clearly, we computed the average numbers of miRNA genes per gene family in all ancestral and extant species (fig. 4). The average number was 2.7 in the ancestor of land plant species, but it gradually increased during the plant evolution examined here (fig. 4). The most extreme case is that of the soybean lineage, where the extant species on average has 8.1 miRNA genes in each gene family. Consequently, the proportion of gene gains by gene duplication is generally high (>60%) in most of the lineages. We also found that more

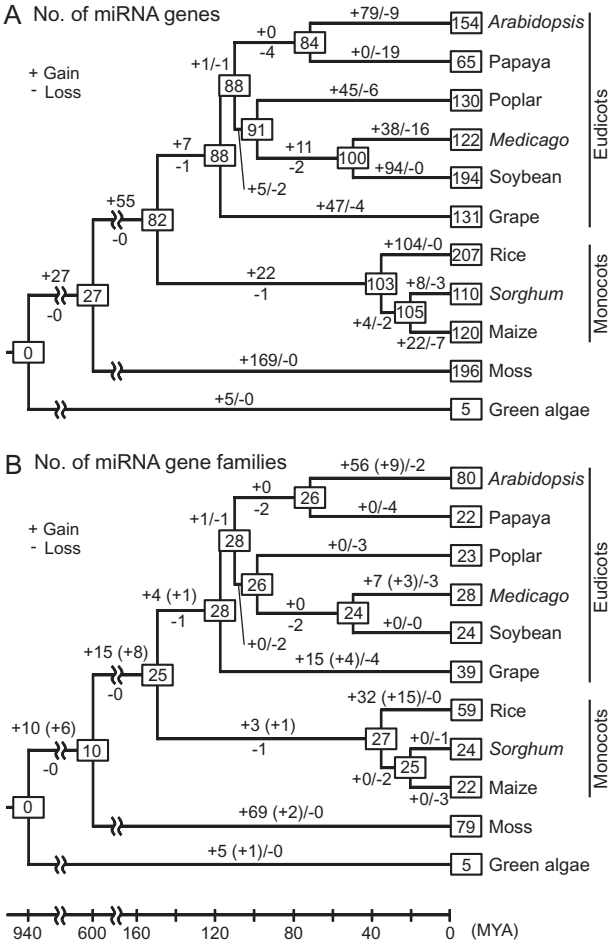

Fig. 3.—Estimates of the numbers of (*A*) miRNA genes and (*B*) gene families in ancestral species and gains and losses of miRNA genes (or gene families) during plant evolution. Numbers in squares represent the numbers of miRNA genes (or gene families) in ancestral or extant species. Numbers along each branch indicate the numbers of gains (+) and losses (−) of miRNA genes (or gene families), respectively. Numbers (+) in parentheses represent the numbers of gains of miRNA gene families that have potentially been derived from protein-coding genes. The time scale shown below the tree is from Hedges and Kumar (2009).

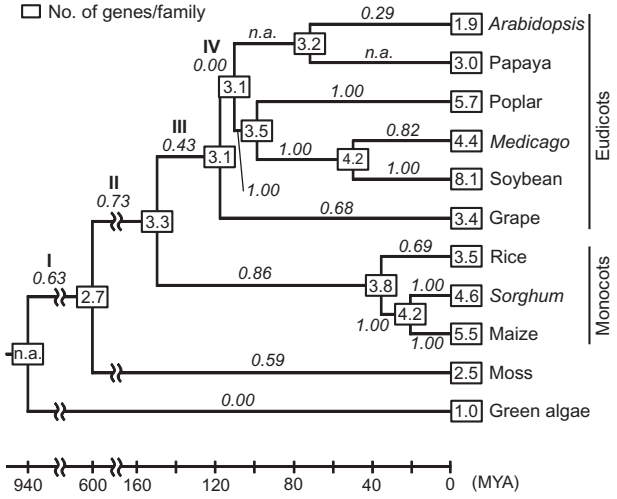

Fig. 4.—Average numbers of miRNA genes per gene family in ancestral or extant species (in squares) and proportions of gene gains by miRNA gene duplication (along each branch in italics). "n.a." means that no gene existed in the ancestor or no gene gain occurred in the lineages. Roman numerals above branches correspond to those in figure 5.

than 90% of gene families which are conserved in nine or more species are multigene families, whereas only 23% of species-specific miRNA genes form multigene families (supplementary fig. S5, Supplementary Material online). These observations suggest that gene duplication has played important roles in increasing the number of miRNA genes in plants.

To obtain some insight into the pattern of miRNA gene losses, we computed the proportions of gene losses that have resulted in gene family losses. The result showed that 40% of the gene losses (31 of 77) during plant evolution have also caused gene family losses. This proportion is significantly greater than the expected value (13%) under the random losses of miRNA genes ($P = 1.2 \times 10^{-12}$ by χ^2 test). This suggests that when an miRNA gene has no paralog, the gene tends to be lost, but once the gene gained an important function, all duplicates of the gene are likely to be maintained in the genome. Because the single miRNA genes are generally young (e.g., species specific) as mentioned above, it appears that young miRNA genes generally have less important functions, whereas old miRNA genes forming multigene families tend to have essential functions. Nevertheless, note that many old miRNA genes can also be lost during long-term plant evolution. For example, five of the ten gene families in the ancestor of land plants and 13 of the 25 gene families in the ancestor of flowering plants were lost in at least one species examined after origination (supplementary table S7, Supplementary Material online). These observations indicate that the repertoires of miRNA genes have dynamically changed in a lineage-specific manner.

miRNA Genes from Protein-Coding Genes

To examine whether some miRNA genes have originated from protein-coding genes, we conducted a homology search (a BLASTN search with a cutoff *E* value of 10^{-5}) by using miRNA genes as queries against all annotated protein-coding genes in the 11 plant species. After removing all miRNA sequences which are located in protein-coding genes, we found 54 miRNA genes that have sequence similarity to CDS of protein-coding genes beyond mature and star regions (table 1). Therefore, these miRNA genes appear to have originated from duplication of protein-coding genes with subsequent mutations. Because these genes belong to 24 different gene families, the proportion of miRNA gene families which seem to have been derived from protein-coding genes was estimated to be ~11% (24/226). If we included the miRNA genes that are homologous to untranslated regions and

Table 1
Numbers of miRNA Genes (gene families), Which are Homologous to Protein-Coding Genes

	Number of miRNA Genes (gene families)			
Species	**CDS**	**UTR**	**Intron**	**Total**
Arabidopsis	9 (9)	0 (0)	9 (2)	18 (11)[a]
Papaya	2 (2)	0 (0)	0 (0)	2 (2)
Poplar	0 (0)	1 (1)	1 (1)	2 (2)
Medicago	3 (3)	19 (3)	8 (2)	30 (7)
Soybean	3 (2)	3 (1)	5 (5)	11 (7)
Grape	20 (4)	1 (1)	4 (3)	25 (8)
Rice	6 (5)	11 (5)	33 (11)	50 (19)
Sorghum	7 (5)	3 (3)	2 (2)	12 (6)
Maize	3 (2)	1 (1)	0 (0)	4 (3)
Moss	0 (0)	0 (0)	2 (2)	2 (2)
Green algae	1 (1)	0 (0)	0 (0)	1 (1)
All	54 (24)	39 (9)	64 (24)	157 (50)

NOTE.—UTR, untranslated region.

[a] Different miRNA genes belonging to the same gene family can have sequence similarity to different regions of protein-coding genes (e.g., CDS and introns). Therefore, total number of gene families can be smaller than the sum of the numbers of gene families for all regions.

introns, the proportion became as large as ~22% (50/226). Note that we did not distinguish miRNA genes derived from simple duplicates of protein-coding genes and those from inverted duplicates of protein-coding genes. It should also be mentioned that the actual proportion of miRNA gene families from protein-coding genes could be even greater because the miRNA genes may have lost the sequence similarity to the protein-coding genes due to subsequent mutations.

We then examined the timing of origination of the miRNA gene families that have potentially been generated from protein-coding genes (fig. 3*B*). The results showed that six of the ten (60%) gene family gains in the ancestral lineage of land plants seemed to have been caused by duplication of protein-coding genes. Similarly, 8 of the 15 (53%) miRNA gene families that originated in the ancestral lineage of flowering plants appeared to have been derived from protein-coding genes. These results imply that this process was important in the early stages of land plant evolution for generating miRNA gene families.

Rates and Patterns of Nucleotide Substitution

To clarify how miRNA genes have evolved after birth, we estimated the rates of nucleotide substitution in the mature and star regions of miRNA genes. We separately computed the rates of nucleotide substitution in miRNA genes, which originated in several different branches leading to *Arabidopsis* (branches I to IV in fig. 4) to examine the relationships between the substitution rate and the time after the birth of miRNA genes (for details, see Materials and Methods). The results showed that there is no clear relationship between the substitution rate and the time after origination (fig. 5). Mature regions showed the rates of ~0.2 × 10^{-9}/site/year, which

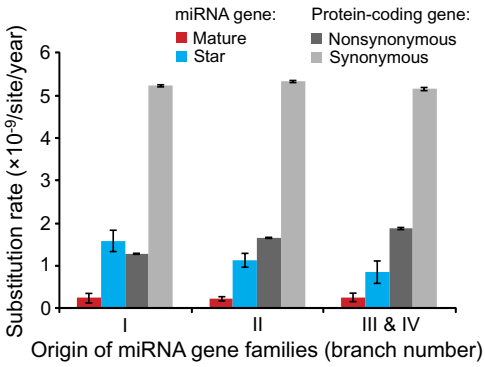

FIG. 5.—Substitution rates of miRNA and protein-coding genes that originated in each branch (I–IV in fig. 4). We analyzed 24 miRNA gene families and 7,147 orthologous protein-coding genes, respectively. Error bars indicate the standard errors. The numbers of miRNA gene families analyzed for each branch are as follows: 7 for branch I, 13 for branch II, and 4 for branches III and IV.

are much slower than the rates for star regions (1–1.5 × 10^{-9}/site/year). This is reasonable because mature sequences are essential to recognize target genes as well as to make a duplex structure with star sequences, whereas star sequences appear to be essential only for keeping the duplex structure with mature sequences. However, even the star sequences showed the rates considerably lower than r_S in protein-coding genes. Therefore, both mature and star regions of miRNA genes have apparently been under purifying selection.

To examine whether some parts of mature and star regions tend to change more frequently than other parts, we calculated the proportion of conserved sites of mature and star regions in the sequence alignments of each miRNA gene family. In this analysis, we subdivided the mature and star regions into three parts. The results showed that upper (5′) portion of the mature region tend to be more conserved than lower (3′) portion (fig. 6*A*). This observation implies that the 5′ portion of the mature region is more important for target recognition and/or miRNA biogenesis than the 3′ portion. By contrast, the middle portion of the star region shows much lower proportion of conserved sites compared with other portions (fig. 6*B*). Consequently, the proportion of paired sites between mature and star regions is significantly lower in the middle portion than in upper and lower portions (supplementary fig. S6, Supplementary Material online). Therefore, although the importance of the small bulge in the middle stem for miRNA biogenesis has been suggested to be small in plants than in animals (Song et al. 2010), some unpaired sites in the middle portion of the duplex structure appear to have some functions even in plants when long-term evolution is considered.

Discussion

In this study, we have examined the evolutionary changes of plant miRNA genes. Our estimation of the numbers of

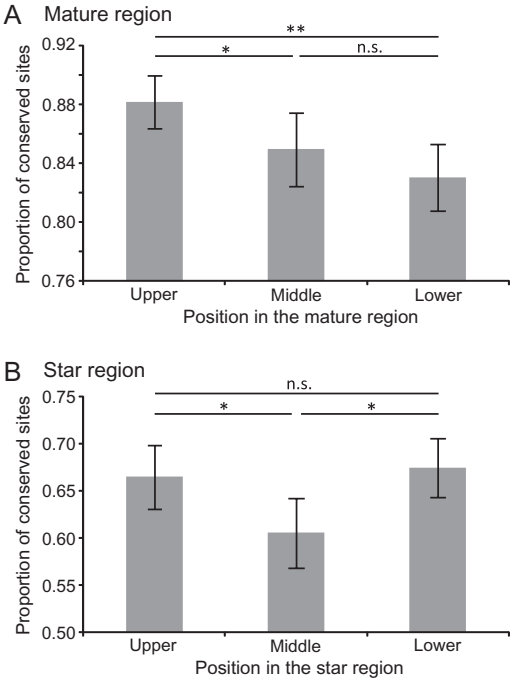

Fig. 6.—Proportions of conserved sites in (*A*) mature and (*B*) star regions. Conserved sites refer to the sites where all sequences have the same nucleotides in the sequence alignments of an miRNA gene family. Each of mature and star regions was equally split into three portions, that is, upper (5′), middle, and lower (3′) portions. Error bars indicate the standard errors. The difference between the portions was tested by Monte Carlo simulation with 10,000 replications: *5% significance level; **1% significance level; n.s., not significant.

miRNA genes and gene families in ancestral species indicated that the numbers of miRNA genes and gene families increased considerably in the lineage to flowering plants after the divergence from green algae. However, these numbers have changed in a lineage-specific manner after the divergence of eudicots and monocots. In fact, more than half of gene families in the ancestors of flowering plants have been lost in at least one species examined. By contrast, 118 gene families have been generated during the evolutionary time for flowering plants (fig. 3*B*). These observations suggest that many old miRNA genes have been lost during plant evolution, and the repertoires of miRNA genes have changed more dynamically in a lineage- or a species-specific manner than previously thought (Zhang et al. 2006; Cuperus et al. 2011). These changes in repertoires of miRNA genes may have affected phenotypic evolution of flowering plants, although this idea is speculative.

For the origin of miRNA genes in plants, we have clearly shown that many miRNA genes have been generated by duplication of preexisting miRNA genes (I in fig. 7). Note that whole-genome duplication (WGD) in addition to tandem duplication of miRNA genes is also included in this process. Indeed, WGD is likely to account for a majority of miRNA gene family expansions in maize compared with *Sorghum* (Zhang et al. 2009). *Arabidopsis* also experienced two rounds of WGD after splitting from papaya (Freeling 2009). In this case, however, the average number of miRNA genes per gene family in *Arabidipsis* is smaller than that in papaya (fig. 4), which is unexpected if WGD generated a large number of miRNA genes. Therefore, the contribution of WGD to the origin of currently existing miRNA genes appears to vary from species to species. Further studies are necessary to clarify this point.

We have also shown that many miRNA gene families seem to have originated from protein-coding genes (II in fig. 7), as was proposed previously (Allen et al. 2004; Rajagopalan et al. 2006; Fahlgren et al. 2007, 2010). This mechanism appears to have been important at the early stage of land plant evolution for generating new miRNA gene families. TEs also seem to be important for generating species-specific miRNA genes (III in fig. 7). In fact, the estimates of the numbers of miRNA genes in grape and rice became considerably greater when we included the TE-like sequences. However, how many of these putative TE-like miRNA genes are functionally relevant remains unclear. Further experimental validation is needed.

In this study, we have not really examined the possibility of miRNA genes arising from hairpin structures in the genome. However, a majority of miRNA gene origination appear to be explained by other mechanisms (I–III in fig. 7). For example, as many as 107 (63%) of the 169 gene gains observed in the lineage to *Arabidopsis* after the divergence from green algae can be explained either by duplication of miRNA genes or protein-coding genes. The proportion becomes even greater in other lineages. Therefore, genomic hairpin structures are unlikely to be a major source of new miRNA genes (IV in fig. 7).

The mechanisms of the origins of miRNA genes in plants are quite different from those in *Drosophila* species, in which the genomic hairpin structures appear to be the major source and the duplication of preexisting miRNA genes is of secondary importance. The contribution of protein-coding genes and TEs seems to be negligible in *Drosophila* species (fig. 7) (Lu et al. 2008; Nozawa et al. 2010; see also Berezikov et al. 2011). The trend seems to be essentially the same in entire evolution of animals (Hertel et al. 2006; see also Sempere et al. 2006; Grimson et al. 2008), although the data set used is quite limited. The reason for this difference may be due to the difference in target recognition between animals and plants. In animals, the seed sequence consisting of only 6–8 nt seems essential for target recognition (Bartel 2009), and therefore, genomic hairpin structures are more likely to generate seed sequences by chance. This is quite plausible because there are hundreds of thousands of hairpin structures in a genome. By contrast, the entire mature sequence (~21 nt) seems to be necessary for binding target sites in plants (Axtell and Bowman 2008), and therefore it would be difficult to generate miRNA genes from hairpin structures in a genome. This hypothesis has been implied

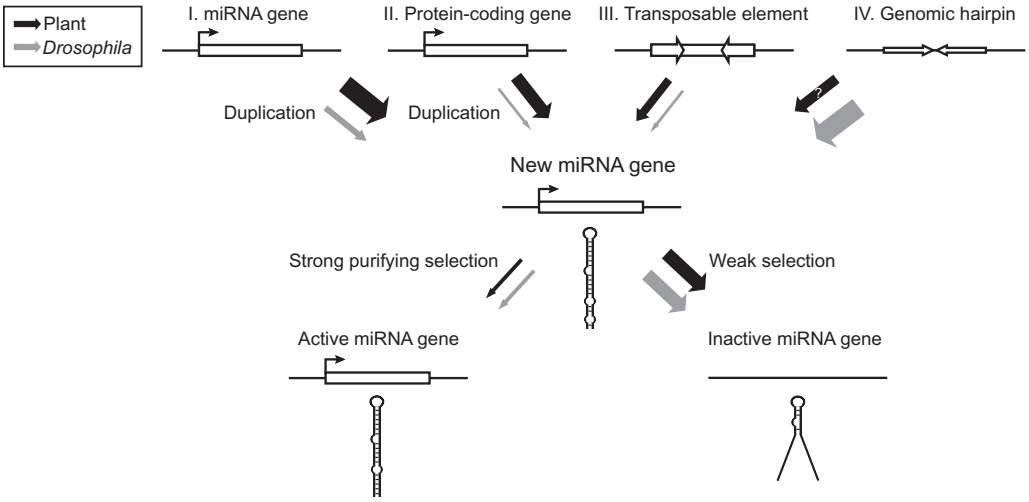

Fig. 7.—Possible evolutionary scenario of miRNA genes in plants (black arrows) in comparison with that in *Drosophila* species (gray arrows). Thickness of arrows roughly indicates the frequencies of the processes.

by several studies (e.g., Li and Mao 2007; Shabalina and Koonin 2008; Voinnet 2009), and our study clearly supports this hypothesis.

We did not find any clear-cut correlation between the evolutionary rate and the time after the birth of miRNA genes in plants. One might think that this observation is contradictory with the previous studies, in which young miRNA genes evolve much faster than old miRNA genes (Fahlgren et al. 2010; Ma et al. 2010). In the present study, however, the youngest miRNA genes analyzed for estimating substitution rates were generated more than 100 Ma (branch IV in fig. 4) and are apparently old enough to have solid functions. In addition, we found that newly generated miRNA genes without any paralogs have been lost more often than the old genes belonging to multigene families. Therefore, our observations together with the findings by previous studies suggest that many young genes are functionally unimportant and only a few of them acquire the solid functions (fig. 7). This pattern of birth-and-death evolution is quite similar to that in *Drosophila* species (Lu et al. 2008; Nozawa et al. 2010).

In this study, we did not examine how the target sites of miRNA genes have evolved in relation to miRNA genes. In general, it can be hypothesized that miRNA genes evolve in a more conservative manner than their target sites. This is because miRNA genes generally regulate the expression of more than one gene (Grun et al. 2005; Lim et al. 2005; Archak and Nagaraju 2007). Consequently, changes in miRNA genes may affect the expression level of several genes concurrently, which could be deleterious to an individual. By contrast, changes in target sites may only affect the target gene, so that the phenotypic effect of the changes could be minor. In fact, miRNA genes in the Hox gene cluster seem to have evolved slower than their target sites in insects (Miura et al. 2011). It would be interesting to study whether these evolutionary patterns of miRNA genes and their target sites can be generalized in animals and plants.

In short, we have shown that the potential sources of miRNA genes are quite different between plants and *Drosophila* species, but the evolutionary pattern of miRNA genes after their originations is similar between them. One might think that because experimental identification of miRNA genes in plants is still limited, our conclusions could be biased. It is certainly quite possible that some unknown miRNA genes are discovered in the future. However, note that the miRNA genes which were used for our homology search were derived from a wide range of plant species including eudicots, monocots, moss, and green algae (supplementary table S2, Supplementary Material online). In addition, even when we used only *Arabidopsis*, rice, and moss, where the extensive experimental data are available, the results were essentially the same (data not shown). Therefore, we believe that our conclusions are robust.

Supplementary Material

Supplementary figures S1–S6, tables S1–S7, miRNA_hairpin.fas, and miRNA_mature.fas are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Mike Axtell, Yuji Iwata, Hielim Kim, Zhaorong Ma, Naoko Takezaki, Liang Song, Yoshiyuki Suzuki, and Zhenguo Zhang for their comments on earlier versions of the manuscript. We also thank Takeshi Itoh and Yoshihiro Kawahara for their advices on our data analysis. This work was supported by National Institutes of Health grant GM020293 to M. Nei and Japan Society for Promotion of Science to M. Nozawa.

Literature Cited

Allen E, et al. 2004. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. Nat Genet. 36:1282–1290.

Archak S, Nagaraju J. 2007. Computational prediction of rice (*Oryza sativa*) miRNA targets. Genomics Proteomics Bioinformatics. 5:196–206.

Axtell MJ. 2008. Evolution of microRNAs and their targets: are all microRNAs biologically relevant? Biochim Biophys Acta. 1779:725–734.

Axtell MJ, Bartel DP. 2005. Antiquity of microRNAs and their targets in land plants. Plant Cell 17:1658–1673.

Axtell MJ, Bowman JL. 2008. Evolution of plant microRNAs and their targets. Trends Plant Sci. 13:343–349.

Axtell MJ, Snyder JA, Bartel DP. 2007. Common functions for diverse small RNAs of land plants. Plant Cell 19:1750–1769.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. Cell 136:215–233.

Berezikov E, et al. 2011. Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. Genome Res. 21:203–215.

Carthew RW, Sontheimer EJ. 2009. Origins and mechanisms of miRNAs and siRNAs. Cell 136:642–655.

Cuperus JT, Fahlgren N, Carrington JC. 2011. Evolution and functional diversification of MIRNA genes. Plant Cell 23:431–442.

De Felippes FF, Schneeberger K, Dezulian T, Huson DH, Weigel D. 2008. Evolution of *Arabidopsis thaliana* microRNAs from random sequences. RNA 14:2455–2459.

Dhandapani V, et al. Forthcoming 2011. Identification of potential microRNAs and their targets in *Brassica rapa* L. Mol Cells. 32:21–37.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Fahlgren N, et al. 2007. High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of *MIRNA* genes. PLoS One. 2:e219.

Fahlgren N, et al. 2010. MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. Plant Cell 22:1074–1089.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 60:433–453.

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool. 28:132–163.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res. 36:D154–D158.

Grimson A, et al. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature 455:1193–1197.

Grun D, Wang YL, Langenberger D, Gunsalus KC, Rajewsky N. 2005. microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. PLoS Comput Biol. 1:e13.

Hedges SB, Kumar S. 2009. The timetree of life. New York: Oxford University Press.

Hertel J, et al. 2006. The expansion of the metazoan microRNA repertoire. BMC Genomics 7:25.

Jones-Rhoades MW, Bartel DP. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell. 14:787–799.

Joshi T, et al. 2010. Prediction of novel miRNAs and associated target genes in *Glycine max*. BMC Bioinformatics 11(1 Suppl):S14.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HM, editor. Mammalian protein metabolism. New York: Academic. p. 21–132.

Klevebring D, et al. 2009. Genome-wide profiling of Populus small RNAs. BMC Genomics 10:620.

Li A, Mao L. 2007. Evolution of plant microRNA gene families. Cell Res. 17:212–218.

Lim LP, et al. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433:769–773.

Lu J, et al. 2008. The birth and death of microRNA genes in *Drosophila*. Nat Genet. 40:351–355.

Ma Z, Coruh C, Axtell MJ. 2010. *Arabidopsis lyrata* small RNAs: transient MIRNA and small interfering RNA loci within the *Arabidopsis* genus. Plant Cell 22:1090–1103.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 288:911–940.

Meyers BC, et al. 2008. Criteria for annotation of plant microRNAs. Plant Cell 20:3186–3190.

Miura S, Nozawa M, Nei M. 2011. Evolutionary changes of the target sites of two microRNAs encoded in the Hox gene cluster of *Drosophila* and other insect species. Genome Biol Evol. 3:129–139.

Morton BR, Bi IV, McMullen MD, Gaut BS. 2006. Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. Genetics 172:569–577.

Nam J, Nei M. 2005. Evolutionary change of the numbers of homeobox genes in bilateral animals. Mol Biol Evol. 22:2386–2394.

Niimura Y, Nei M. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One. 2:e708.

Nozawa M, Miura S, Nei M. 2010. Origins and evolution of microRNA genes in *Drosophila* species. Genome Biol Evol. 2:180–189.

Oki N, et al. 2008. A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. *japonica*. Genes Genet Syst. 83:321–329.

Piriyapongsa J, Jordan IK. 2008. Dual coding of siRNAs and miRNAs by plant transposable elements. RNA 14:814–821.

Rajagopalan R, Vaucheret H, Trejo J, Bartel DP. 2006. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. Genes Dev. 20:3407–3425.

Sempere LF, Cole CN, McPeek MA, Peterson KJ. 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. J Exp Zoolog B Mol Dev Evol. 306:575–588.

Shabalina SA, Koonin EV. 2008. Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol. 23:578–587.

Song L, Axtell MJ, Fedoroff NV. 2010. RNA secondary structural determinants of miRNA precursor processing in *Arabidopsis*. Curr Biol. 20:37–41.

Voinnet O. 2009. Origin, biogenesis, and activity of plant microRNAs. Cell 136:669–687.

Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA. 2006. Conservation and divergence of plant microRNA genes. Plant J. 46:243–259.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A. 95:3708–3713.

Zhang L, et al. 2009. A genome-wide characterization of microRNA genes in maize. PLoS Genet. 5:e1000716.

Zhu QH, et al. 2008. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. Genome Res. 18:1456–1465.

Associate editor: Marta Wayne