Origin and Evolutionary Dynamics of the miR2119 and ADH1 Regulatory Module in Legumes

Carlos De la Rosa[1,4,*], Luis Lozano[2,3], Santiago Castillo-Ramírez[2,3], Alejandra A. Covarrubias[1], and José L. Reyes[1,*]

[1]Departamento de Biología Molecular de Plantas, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

[2]Luis Lozano Unidad de Análisis Bioinformáticos, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de Mexico, Cuernavaca, México

[3]Santiago Castillo Programa de Genómica Evolutiva, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de Mexico, Cuernavaca, Mexico

[4]Present address: Departamento de Investigaciones Científicas y Tecnológicas, Universidad de Sonora, Blvd. Luis D. Colosio S/N entre Reforma y Sahuaripa, Col Centro, Hermosillo, Mexico

*Corresponding authors: E-mails: carlos.delarosa@unison.mx; jlreyes@ibt.unam.mx.

Accepted: 29 September 2020

Abstract

MicroRNAs are important regulators of gene expression in eukaryotes. Previously, we reported that in *Phaseolus vulgaris*, the precursor for miR2119 is located in the same gene as miR398a, conceiving a dicistronic *MIR* gene. Both miRNA precursors are transcribed and processed from a single transcript resulting in two mature microRNAs that regulate the mRNAs encoding ALCOHOL DEHYDROGENASE 1 (ADH1) and COPPER-ZINC SUPEROXIDE DISMUTASE 1 (CSD1). Genes for miR398 are distributed throughout the spermatophytes; however, miR2119 is only found in Leguminosae species, indicating its recent emergence. Here, we used public databases to explore the presence of the miR2119 sequence in several plant species. We found that miR2119 is present only in specific clades within the Papilionoideae subfamily, including important crops used for human consumption and forage. Within this subfamily, *MIR*2119 and *MIR*398a are found together as a single gene in the genomes of the Millettioids and Hologalegina. In contrast, in the Dalbergioids *MIR*2119 is located in a different locus from *MIR*398a, suggesting this as the ancestral genomic organization. To our knowledge, this is a unique example where two separate *MIRNA* genes have merged to generate a single polycistronic gene. Phylogenetic analysis of *ADH1* gene sequences in the Papilionoideae subfamily revealed duplication events resulting in up to four *ADH1* genes in certain species. Notably, the presence of *MIR*2119 correlates with the conservation of target sites in particular *ADH1* genes in each clade. Our results suggest that post-transcriptional regulation of *ADH1* genes by miR2119 has contributed to shaping the expansion and divergence of this gene family in the Papilionoideae. Future experimental work on *ADH1* regulation by miR2119 in more legume species will help to further understand the evolutionary history of the *ADH1* gene family and the relevance of miRNA regulation in this process.

Key words: microRNA evolution, dicistronic miRNA precursor, alcohol dehydrogenase 1, miR398.

Significance

The plant microRNA miR2119 is present only in specific clades within the Papilionoideae subfamily, including important crops used for human consumption and forage. In some species, miR2119 is processed from a dicistronic transcript also containing miR398a to regulate the expression of ADH1 and CSD1 transcripts, respectively. Here we performed an exploration of different plant genome and small RNA databases to study the prevalence of the miR2119 precursor and the *ADH1* target genes. Our results indicate that several genomic rearrangement events have occurred, shaping the genomic organization of *MIR*2119 and that of its corresponding target *ADH1* genes.

Introduction

Legumes (Leguminosae or Fabaceae) are the third-largest plant family with around 20,000 species. Grains derived from legumes provide one-third of the protein in the human diet and also contribute to about a third of vegetable oil used for human consumption. In addition, legumes are also important for the production of temperate-climate forage species (alfalfa, *Trifolium pratense*) or tropical climate species (*Stylosanthes*, *Desmodium*) (Graham and Vance 2003; Gepts et al. 2005).

The legume family maintains a cosmopolitan distribution, representing an important ecological constituent and has a widespread use in agricultural systems. Although not all legumes form an association with nitrogen-fixing bacteria (Griesmann et al. 2018), the ability of most legume species to fix nitrogen through symbiosis with bacteria from the genus *Rhizobium* is perhaps one of the best-known features of this family. Bacteria can convert atmospheric nitrogen into ammonium by the enzyme nitrogenase, this process occurs inside specialized organs in the root called nodules. The nitrogen fixed is ceded to the host plant for use in the synthesis of essential compounds such as amino acids, nucleic acids, among others (Dos Santos et al. 2012). In general, the legume family is exceptionally diverse in morphology, physiology, and in ecological terms; thus, this family represents one of the most interesting known examples in evolutionary aspects and diversification in plants (Azani et al. 2017).

Recently, an international community studying legumes systematics classified the legume family into six subfamilies: Caesalpinioideae (including clade Mimosoideae), Cercidoideae, Detarioideae, Dialioideae, Duparquetioideae, and Papilionoideae (Azani et al. 2017). This classification was based on a phylogenetic analysis of the plastid gene *matK* sequence, which included almost all the genera (698 of the 765 recognized genera) and ∼20% of the species (3,696) known to date. This novel classification is the most complete evolutionary study of legumes known thus far (Azani et al. 2017). In particular, the Papilionoideae subfamily contains legumes that provide food and are economically important to human beings (Doyle and Luckow 2003). As part of the Papilionoideae subfamily, there are four important clades Genistoids, Dalbergioids, Hologalegina, and Millettioids (Gepts et al. 2005). The Genistoids clade includes the genus *Lupinus* and the Dalbergioids clade contains the genera *Arachis* and *Nissolia* represented by *Arachis hypogaea* (peanut) and *Nissolia schottii*. The Hologalegina clade is divided into two subclades: Robinioids represented by *Lotus japonicus*, and IRLC (for its acronym *Inverted Repeat Lacking Clade*), which includes species characterized by the loss of a copy of an inverted repeat in the chloroplast DNA found in most angiosperms. The IRLC subclade includes species such as *Medicago sativa* (alfalfa), *Cicer arietinum* (chickpea), *Vicia faba* (faba bean), *Lens culinaris* (lentil), and *Pisum sativum* (pea). Finally, the Millettioids clade includes several legumes that are better adapted to tropical climates and, therefore, were named as warm season legumes, including *Phaseolus vulgaris* (common bean), *Vigna unguiculata* (cowpea), *Cajanus cajan* (pea bean or pigeon pea), and *Glycine max* (soybean) (Doyle and Luckow 2003; Gepts et al. 2005). Representative clades in the Papilionoideae subfamily can be seen in figure 1.

MicroRNAs (miRNAs) are important regulators of gene expression at the post-transcriptional level in animals and plants. These small RNA molecules are generated from a double-stranded precursor by the action of DICER-LIKE 1 (DCL1), an RNAse III family endonuclease that produces mature miRNAs about 21–22 nt in length. In complex with an Argonaute protein, miRNAs catalyze the recognition of target mRNAs through base-pairing resulting in the inhibition of their expression by RNA cleavage or translation inhibition (Axtell 2013). In plants, conserved miRNAs are present in non-vascular and vascular plants. Within individual plant families, less-conserved miRNAs regulate family-specific processes, relevant for their own lifestyles. We have previously shown that in common bean, miR2119 regulates the expression of *ADH1* in response to water deficit, and that *MIR*2119 is encoded in a dicistronic transcript together with *MIR*398a, which is a different miRNA targeting the transcript for CSD1 (De la Rosa et al. 2019). We reported the function of miR2119 in *P. vulgaris* and also provided evidence for its presence in other legumes such as *G. max*, *Medicago truncatula*, and *A. hypogaea* (Arenas-Huertero et al. 2009; De la Rosa et al. 2019). To expand our analysis on the distribution of the *MIR*398-*MIR*2119 gene, we carried out an exploration in different plant genome databases to study the prevalence of this precursor and the *ADH1* target genes. Our results indicate that within the Papilionoideae subfamily several genomic rearrangement events have shaped the current genomic organization of *MIR*2119 and its target *ADH1* genes; thus, likely affecting the patterns of mRNA regulation within the *ADH1*-*MIR*2119 module.

Materials and Methods

Databases Used

We explored genome sequences available from different legumes in databases including NCBI (www.ncbi.nlm.nih.gov): *Phaseolus coccineus* UCLA_Phcoc_1.0, *Glycine soja* ASM419377v2, *Cicer reticulatum* ASM368901v2, *Cicer echinospermum* S2Drd065_v0.5, *Trifolium medium* ASM349008v1, *Trifolium subterraneum* TSUd_r1.1, *P. sativum* ASM301357v1, *Arachis monticola* ASM306328v2, *N. schottii* ASM325490v1, *Mimosa pudica* ASM325494v1 and *Cercis canadensis* ASM325506v1; in the Phytozome database (phytozome.jgi.doe.gov): *P. vulgaris* v2.1, *G. max* Wm82.a2.v1, and *M. truncatula* Mt4.0v1; in the Legume

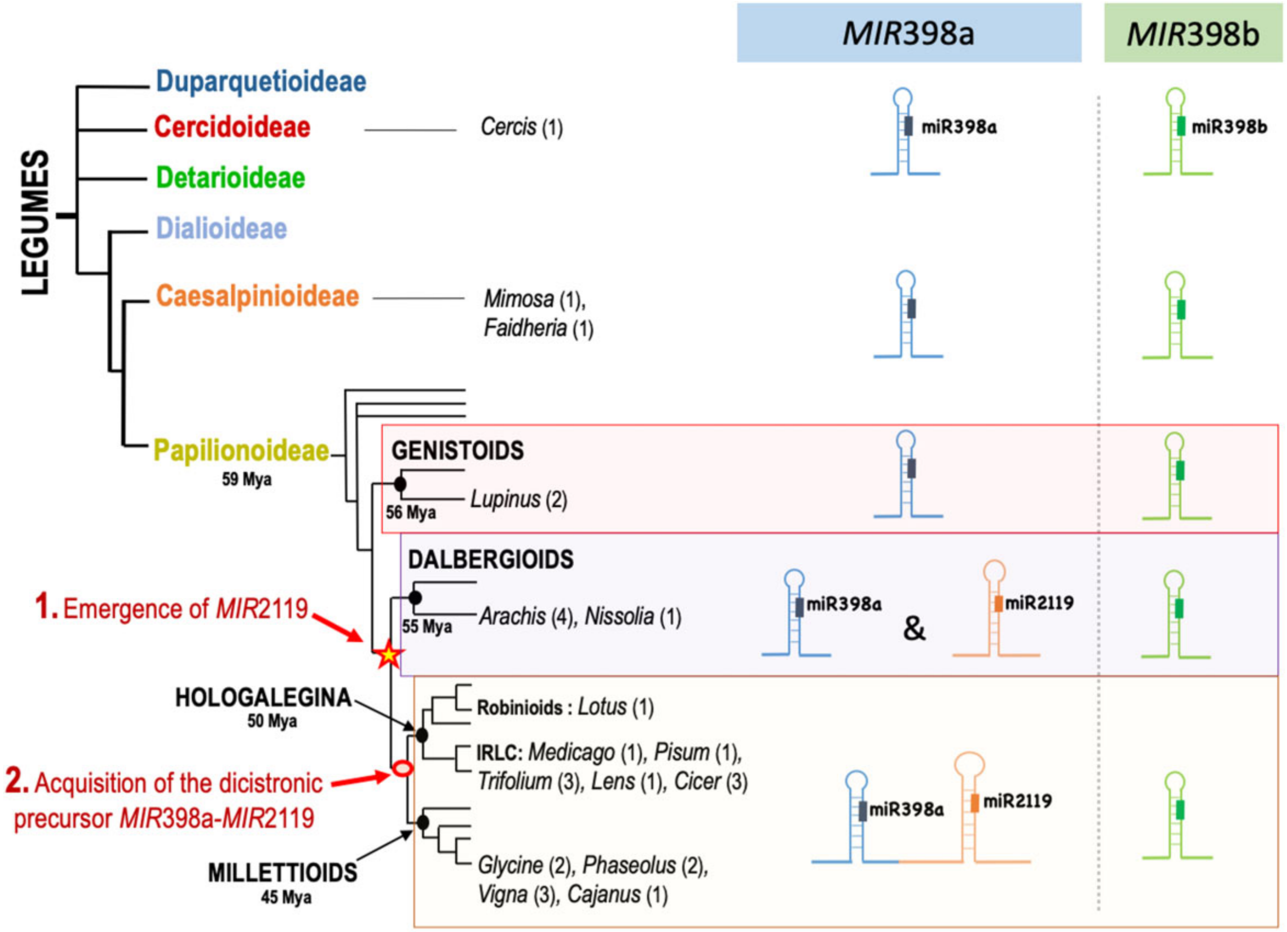


Fig. 1.—Emergence of *MIR*2119 and acquisition of the dicistronic *MIR*398a–*MIR*2119 gene in the Papilionoideae subfamily. In the Papilionoideae subfamily of legumes, there are four important clades: Genistoids, Dalbergioids, Hologalegina, and Millettioids. The star symbol indicates the suggested point of emergence of *MIR*2119 among the common ancestor of the Dalbergioids and Hologalegina-Millettioids clades. The circle in red indicates the acquisition of the dicistronic *MIR*398a–*MIR*2119 gene, which likely arose in the common ancestor of the Hologalegina and Millettioids clades. A number within parentheses indicates genomes analyzed in each genus. The legume family dendrogram was based on Gepts et al. (2005), including the estimated time of divergence (Ma); modified, and updated based on Azani et al. (2017).

Information System database (www.legumeinfo.org/): *Vigna angularis* v3.0, *Vigna radiata* v1.0, *V. unguiculata* IT97K-499-35 v1.0, *C. cajan* v1.0, *L. japonicus* v3.0, *C. arietinum* ICC4958.v2.0, *T. pratense* v2.0, *Arachis duranensis* v1.0, *Arachis ipaensis* v1.0, *A. hypogaea* v1.0, *Lupinus angustifolius* v1.0, *Lupinus albus* v.1.0, and *Faidheria albida* v.1.0; as well as the genome sequence of *L. culinaris* (UofS, v1.2) included in the KnowPulse database (knowpulse.usask.ca/).

miR398 and miR2119 Gene Sequences

The sequences for *P. vulgaris* pre-miR398-miR2119 and pre-miR398b (Chromosome 2 pos.9731038-9732110 and Chromosome 8 pos. 54889992-54890117 negative strand, respectively) were used as queries to identify related sequences using the BLASTN program in the collection of *Expressed Sequence Tags* (ESTs), mRNAs, and genomic sequences in the legumes described above. To expand our search, we used some of the resulting sequences to perform a subsequent BLASTN search and identify more divergent candidate sequences. Each obtained full-length sequence was used to predict its potential secondary structure in search for the fold-back expected for miRNA precursors using the Mfold software (mfold.rna.albany.edu) (Zuker 2003), and then we confirmed the position of the mature miRNA within the stem region.

ADH1 Gene Sequences

The gene sequences for *ADH1.1* (Phvul.009G134700), *ADH1.2* (Phvul.001G064000), *ADH1.3* (Phvul.001G06300), and *ADH1.4* (Phvul.009G149500) of *P. vulgaris* cultivar G19833 were obtained from the Phytozome database. To retrieve other *ADH1* sequences, we first identified a phylogenetic tree of the *ADH1* gene family containing sequences belonging to eight legume and five nonlegume species, available in the *Gene family and phylogenetic tree* section (Dash et al. 2016) of the LIS website (https://legumeinfo.org/, last accessed October 12, 2020). To expand this information, we obtained all ADH1 protein sequences available therein.

Phylogenetic Analyses

The phylogenetic reconstruction of the *ADH1* gene family was made based on 66 protein sequences obtained from the Legume Information System database. Some of the protein sequences were manually curated to correct annotation errors and only those sequences comprising above 90% of the total protein length (average of 380 aa) were selected. The *ADH2* gene (AT5G43940.1) from *Arabidopsis thaliana* was selected as an outgroup for these analyses. ADH2 is a class III ADH also referred to as nitrosoglutathione reductase (Xu et al. 2013). The *ADH2* genes define a separate clade, independent of all other *ADH1* and *ADH1*-like genes present in land plants (Bui et al. 2019). The 67 protein sequences were aligned with the program MUSCLE V3.8.31 (Edgar 2004). Afterwards, we used the ProtTest 3.4.2 program which determined JJT+G as the best-fit substitution model for the alignment (Darriba et al. 2011). The *maximum likelihood* method (ML) phylogeny was built with PhyML 3.0 program with SH-like support values considered as significant if higher than 0.7 (Guindon et al. 2009). The phylogenetic tree was visualized with the program FigTree V1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/, last accessed October 12, 2020). To estimate possible duplication events, we employed the NOTUNG 2.9.1.5 program using default parameters (Stolzer et al. 2012). As species tree we used the ML phylogeny of the *mat*K gene constructed again via PhyML 3.0, setting the model to GTR + I+G, which was the best model as per jModelTest (Posada 2009).

Other Bioinformatical Tools Used

The RNAhybrid program (Kruger and Rehmsmeier 2006) was used to determine and calculate the most favorable hybridization site between each *ADH1* gene sequence and the corresponding miR2119 sequence for each species analyzed. For prediction of the consensus sequences and sequence alignments, we employed the Meme suite 5.0.4, Clustal-O program (Bailey et al. 2009; Sievers et al. 2011) and the T-coffee program (Notredame et al. 2000; Di Tommaso et al. 2011).

Results

miR2119 Is Present Only in Specific Clades within the Papilionoideae Subfamily

In order to identify potential homologous sequences for miR2119 in other legume species, we first conducted BLAST searches, using the miR398a-miR2119 and miR398b precursors of *P. vulgaris* as queries, against the ESTs, mRNAs, and genomic sequences in the genomes of legumes present in NCBI, Phytozome, the Legume Information System (LIS), and KnowPulse databases. To expand this approach, we also employed some of the obtained sequences in subsequent BLAST searches to uncover more divergent sequences.

The sequence data obtained for the mature sequence of miR398 and miR2119 in legumes are summarized in tables 1 and 2, respectively. Each of the identified precursor miRNA sequences was subjected to an in silico secondary structure prediction using the Mfold program using default parameters (Zuker 2003). Most sequences conformed to the expected structure for miRNA precursors with the exception of some isoforms of miR398 in the genus *Arachis* such as miR398b of *A. duranensis* and *A. ipaensis*, miR398d and miR398e in *A. hypogaea*, and miR398c and miR398d in *A. monticola.* Their predicted secondary structure showed limited complementarity in the stem region due to the presence of nine consecutive adenosine residues upstream of the mature miRNA, which reduces the stability of the secondary structure; however, it is likely that this array of adenosines is present due to sequencing or assembly errors. Despite this, the mature sequences of these isoforms were retained for further analysis because of their high identity to the canonical miR398a sequence.

Our previous analysis of the *P. vulgaris*, *G. max*, and *M. truncatula* genomes revealed two kinds of *MIR*398 loci: one where the transcript contains the precursors for miR398 and miR2119, and another where *MIR*398 remains as an independent transcriptional unit and is similar to the loci found in species outside legumes (De la Rosa et al. 2019). In *A. thaliana*, there are three loci for the *MIR*398 gene family: *MIR*398a, *MIR*398b, and *MIR*398c, whereas *Oryza sativa* (rice) contains two loci encoding *MIR*398a and *MIR*398b (Jones-Rhoades and Bartel 2004; Sunkar and Zhu 2004). Our search for sequences in the different databases revealed that most legume genomes analyzed possess at least two *MIR*398 loci, whereas the genomes of *G. max* and of *A. hypogaea* contain six and five loci for *MIR*398, respectively. In addition, for *P. vulgaris*, we identified another locus for *MIR*398 in chromosome 6, named here as *MIR*398c, whose mature miRNA differs in four positions from miR398a (table 1). We did not find any potential small RNA in its vicinity, as is the case for the *MIR*398b gene. It was previously described that miR398 is conserved in spermatophytes (Jones-Rhoades and Bartel 2004; Sunkar and Zhu 2004). In particular, the sequence of miR398a is highly conserved and was almost identical in each of the legume species analyzed, indicating that in all cases it regulates the transcript encoding for CSD1 as it has been demonstrated in several plant species (Zhu et al. 2011). Together, these data indicate that the organization of the *MIR*398 gene family in legumes is similar to that of other plant species, except for the presence of MIR2119 in certain loci, as we describe below.

Table 1
miR398 Sequences Identified in Legumes

Organism		Sequence	Mapping	Position	Database
Phaseolus vulgaris	miR398a	UGUGUUCUCAGGUCACCCCUU	Chr02	9731143..9731163	Phytozome
	miR398b	UGUGUUCUCAGGUCGCCCCUG	Chr08	54890009..54890029 (-)	
	miR398c	UGUGUUCUCAGGUCGCUUCUG	Chr06	29983237..29983257 (-)	
Phaseolus coccineus	miR398a	UGUGUUCUCAGGUCACCCCUU	QBDZ01159137	1394..1414 (-)	NCBI
	miR398b	UGUGUUCUCAGGUCGCCCCUG	QBDZ01190595	19025-19045 (-)	
	miR398c	UGUGUUCUCAGGUCGCUCCUC	QBDZ01192480	2117..2137	
Phaseolus acutifolius	miR398a	UGUGUUCUCAGGUCACCCCUU	EST: HO796397	1043..1063 (-)	NCBI
Vigna radiata	miR398a	UGUGUUCUCAGGUCACCCCUU	scaffold_100	976412..976432	LIS
	miR398b	UGUGUUCUCAGGUCGCCCCUG	Vr06	2914498..2914518 (-)	
Vigna angularis	miR398a	UGUGUUCUCAGGUCACCCCUU	vigan.scaffold_5	327943..327963	LIS
	miR398b	UGUGUUCUCAGGUCGCCCCUG	Va01	5116345.. 511636	
Vigna unguiculata	miR398a	UGUGUUCUCAGGUCACCCCUU	Vu02	19512207..19512227	LIS
	miR398b	UGUGUUCUCAUGUCACUUCUU	Vu02	19522073..19522093	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	Vu08	35309954..35309974	
	miR398d	UGUGUUCUCAGGUCGCCCCUG	Vu06	33097218..33097238 (-)	
Glycine max	miR398a	UGUGUUCUCAGGUCACCCCUU	Chr02	11081015..11081035 (-)	Phytozome
	miR398b	UGUGUUCUCAGGUCACCCCUU	Chr01	7214768..7214768 (-)	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	Chr08	14229989..14230009 (-)	
	miR398d	UGUGUUCUCAGGUCGCCCCUG	Chr02	46102437..46102457	
	miR398e	UGUGUUUUCAGGUCACCCAUG	Chr14	2694696..2694716 (-)	
	miR398f	UCUGUUCUCAGGUCGCCCUUG	Chr15	4337756..4337776	
Glycine soja	miR398a	UGUGUUCUCAGGUCACCCCUU	CM009366	7311716..7311736 (-)	NCBI
	miR398b	UGUGUUCUCAGGUCACCCCUU	CM009367	11364601..11364621 (-)	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	CM009373	14536894..14536914 (-)	
	miR398d	UGUGUUCUCAGGUCGCCCCUG	CM009367	48771094..48771114	
	miR398e	UGUGUUUUCAGGUCACCCAUG	CM009379	2816972..2816992 (-)	
	miR398f	UCUGUUCUCAGGUCGCCCUUG	CM009380	4356798..4356818	
Cajanus cajan	miR398a	UGUGUUCUCAGGUCACCCCUU	Cc06	7041889..7041909	LIS
	miR398b	UGUGUUCUCAGGUCGCCCCUG	Cc02	12942141..12942161	
Lotus japonicus	miR398a	UGUGUUCUCAGGUCACCCCUU	Lj0	55824356..55824376 (-)	LIS
	miR398b	UGUGUUCUCAGGUCACCCCUU	Lj0	93079050..93079070	
	miR398c	UGUGUUCUCAGGUCACCCCUU	Lj3	16549333..16549353 (-)	
	miR398d	UGUGUUCUCAGGUCGCCCCUG	Lj2	38990632..38990652	
Cicer arietinum	miR398a	UGUGUUCUCAGGUCACCCCUU	Ca2	22138145..22138165	LIS
	miR398b	UGUGUUCUCAGGUCGCCCCUG	Ca2	4829006..4829026 (-)	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	Ca2	4880660..4880680 (-)	
	miR398d	UGUGUUCUCAGGUCGCCCCUG	Ca2	4742878..4742898	
Cicer reticulatum	miR398a	UGUGUUCUCAGGUCACCCCUU	CM010872	22687187..22687207	NCBI
	miR398b	UGUGUUCUCAGGUCGCCCCUG	CM010872	4053621..4053641	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	CM010872	4137572..4137592 (-)	
Cicer echinospermum	miR398a	UGUGUUCUCAGGUCACCCCUU	PGTU01016578	14915..14935 (-)	NCBI
	miR398b	UGUGUUCUCAGGUCGCCCCUG	PGTU01018136	238749..238769	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	PGTU01018136	321878..321898 (-)	
Medicago truncatula	miR398a	UGUGUUCUCAGGUCACCCCUU	chr5	19181153..19181173 (-)	Phytozome
	miR398b	UGUGUUCUCAGGUCGCCCCUG	chr5	38762041..38762061	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	chr7	3768799..3768819 (-)	
Trifolium pratense	miR398a	UGUGUUCUCAGGUCACCCCUU	Tp57577_LG2	8753507..8753527	Phytozome
	miR398b	UGUGUUCUCAGGUCACCCCUU	Tp57577_LG2	18586621..18586641	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	Tp57577_LG4	2422070..2422090 (-)	
Trifolium medium	miR398b	UGUGUUCUCAGGUCGCCCCUG	LXQA011140102	148..168 (-)	NCBI
Trifolium subterraneum	miR398a	UGUGUUCUCAGGUCACCCCUU	DF973777	105122..105142	NCBI
	miR398b	UGUGUUCUCAGGUCGCCCCUG	DF973242	64770..64790	
Pisum sativum	miR398a	UGUGUUCUCAGGUCACCCCUU	PUCA013739517	14511..14531	NCBI

(continued)

Table 1 Continued

Organism		Sequence	Mapping	Position	Database
	miR398b	UGUGUUCUCAGGUCGCCCCUG	PUCA012795113	19254..19274 (-)	
Lens culinaris	miR398a	UGUGUUCUCAGGUCACCCCUU	LcChr5	55469814..55469834 (-)	KnowPulse
	miR398b	UGUGUUCUCAGGUCGCCCCUG	LcContig611472	11320..11340	
	miR398c	UGUGUUCUCAGGUCGUUCCUG	LcChr3	173110615..173110635 (-)	
Arachis duranensis	miR398a	UGUGUUCUCAGGUCACCCCUU	Aradu.A09	104766867..104766887	LIS
	miR398b	UGUGUUCUCAGGUCGCCCCUG	Aradu.A07	4959034..4959054	
Arachis ipaensis	miR398a	UGUGUUCUCAGGUCACCCCUU	Araip.B09	127447277..127447297 (-)	LIS
	miR398b	UGUGUUCUCAGGUCGCCCCUG	Araip.B03	5114128..5114148	
Arachis hypogaea	miR398a	UGUGUUCUCAGGUCACCCCUU	Arahy.07	57200647..57200667 (-)	LIS
	miR398b	UGUGUUCUCAGGUCACCCCUU	Arahy.09	105785997..105786017 (-)	
	miR398c	UGUGUUCUCAGGUCACCCCUU	Arahy.19	137958044..137958064 (-)	
	miR398d	UGUGUUCUCAGGUCGCCCCUG	Arahy.07	4028484..4028504 (-)	
	miR398e	UGUGUUCUCAGGUCGCCCCUG	Arahy.13	5221799..5221819	
Arachis monticola	miR398a	UGUGUUCUCAGGUCACCCCUU	CM009791	13457618..134576838 (-)	NCBI
	miR398b	UGUGUUCUCAGGUCACCCCUU	CM009781	104209227..104209247 (-)	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	QBTX01000189	114738..114758	
	miR398d	UGUGUUCUCAGGUCGCCCCUG	CM009785	6526804..6526824	
Nissolia schottii	miR398a	UGUGUUCUCAGGUCACCCCUU	QANU01088005	166936..166956	NCBI
	miR398b	UGUGUUCUCAGGUCACCCCUU	QANU01070409	10590..10610 (-)	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	QANU01029087	10731..10751	
Lupinus angustifolius	miR398a	UGUGUUCUCAGGUCACCCCUU	NLL-11	7727180..7727200	LIS
	miR398b	UAUGUUCUCAGGUCGCCCCUG	NLL-09	21047182..21047202 (-)	
Lupinus albus	miR398a	UGUGUUCUCAGGUCACCCCUU	Lalb_Chr10	13947294..13947314 (-)	LIS
	miR398b	UGUGUUCUCAGGUCGCCCCUG	Lalb_Chr10	18348914..18348934	
Mimosa pudica	miR398a	UGUGUUCUCAGGCCACCCCUA	QANV01072731	137075..137095 (-)	NCBI
	miR398b	UGUGUUCUCAGGCCACCCCUA	QANV01054059	5580..5600	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	QANV01051282	29875..29895	
Faidherbia albida	miR398a	UGUGUUCUCAGGUCACCCCUU	scaffold2728_cov186	170829..170849	LIS
	miR398b	UGUGUUCUCAGGUCACCCCUU	scaffold2728_cov186	232576..232596	
	miR398c	UGUGUUCUCAGGUCGCCCCUG	scaffold1096_cov196	330016..330036	
Cercis Canadensis	miR398a	UGUGUUCUCAGGUCACCCCUU	QAOA01003368	343714..343734 (-)	NCBI
	miR398b	UGUGUUCUCAGGUCGCCCCUG	QAOA01003028	484703..484723 (-)	
	miR398c	UAUGUUCUCAGGUCGCCCCUG	QAOA01002999	272469..272489	
Arabidopsis thaliana	miR398a	UGUGUUCUCAGGUCACCCCUU	Chr2	1041012..1041032	Phytozome
	miR398b	UGUGUUCUCAGGUCACCCCUG	Chr5	4691107..4691127	
	miR398c	UGUGUUCUCAGGUCACCCCUG	Chr5	4694778..4694798	
Oryza sativa	miR398a	UGUGUUCUCAGGUCACCCCUU	Chr10	9216260..9216280 (-)	Phytozome
	miR398b	UGUGUUCUCAGGUCGCCCCUG	Chr7	14598627..14598647 (-)	

Note.—The table shows the name of the species, miR398 isoforms and their sequences, the fragment and the position where this sequence is located, and the information source (NCBI, Phytozome, Legumes information System [LIS] or KnowPulse). For each sequence, the position in gray highlights the base change with respect to the *P. vulgaris* miR398a sequence. In mapping, EST, Chr: Chromosome, contig, scaffold, or identifier number indicate assembled sequences or fragments of the genome. In position, (-) indicates the sequence is located in the opposite strand. The version of each database used can be found in the Materials and Methods.

We previously characterized miR2119 as a legume-specific miRNA (Arenas-Huertero et al. 2009; De la Rosa et al. 2019). The results obtained from the search for miR2119 sequences in the available genomes showed its presence only in species belonging to the Papilionoideae subfamily, as detailed in table 2. We identified the sequence of miR2119 in the genome sequences of Millettioids, Hologalegina, and Dalbergioids, but not in the Genistoids. The Millettioids are represented by *P. vulgaris*, *P. coccineus*, *Phaseolus acutifolius*, *V. radiata*, *V. angularis*, *V. unguiculata*, *G. max*, and *G. soja*, and all have an identical miR2119 sequence except for *C. cajan*, which differs in the first position (1C), and *V. unguiculata* that contains an additional copy (miR2119b) with three substitutions (6A, 14C, and 17U). In the Hologalegina clade, there are species belonging to the IRLC subclade such as *M. truncatula*, *T. pratense*, *T. medium*, *T. subterraneum*, *P. sativum*, and *L. culinaris*, which share the same miR2119 sequence; whereas *C. arietinum* and *C. reticulatum* show two changes at positions 9G and 14A. In *L. japonicus* (Robinioids subclade), the miR2119 sequence differs in the second position (2A) with respect to *M. truncatula*. Considering the Dalbergioid clade, species within the genus *Arachis* (*A. duranensis*, *A. ipaensis*, *A. hypogaea*, and *A. monticola*) contain an identical sequence for miR2119, whereas the latter

Table 2
miR2119 Sequences Identified in Legumes

Organism		Sequence	Mapping	Position	Database
Phaseolus vulgaris	miR2119	UCAAAGGGAGUUGUAGGGGAA	Chr02	9731434..9731454	Phytozome
Phaseolus coccineus	miR2119	UCAAAGGGAGUUGUAGGGGAA	QBDZ01159137	1123..1143 (-)	NCBI
Phaseolus acutifolius	miR2119	UCAAAGGGAGUUGUAGGGGAA	HO796397	845..865 (-)	NCBI
Vigna radiata	miR2119	UCAAAGGGAGUUGUAGGGGAA	scaffold_100	976653..976673	LIS
Vigna angularis	miR2119	UCAAAGGGAGUUGUAGGGGAA	vigan.scaffold_5	328184..328204	LIS
Vigna unguiculata	miR2119a	UCAAAGGGAGUUGUAGGGGAA	Vu02	19512408..19512428	LIS
	miR2119b	UCAAAAGGAGUUGCAGUGGAA	Vu02	19522269..19522289	
Glycine max	miR2119a	UCAAAGGGAGUUGUAGGGGAA	Chr02	11080751..11080771 (-)	Phytozome
	miR2119b	UCAAAGGGAGUUGUAGGGGAA	Chr01	7214498..7214518 (-)	
Glycine soja	miR2119a	UCAAAGGGAGUUGUAGGGGAA	CM009366	7311446..7311466 (-)	NCBI
	miR2119b	UCAAAGGGAGUUGUAGGGGAA	CM009367	11364337..11364357(-)	
Cajanus cajan	miR2119	CCAAAGGGAGUUGUAGGGGAA	Cc06	7042140..7042160	LIS
Lotus japonicus	miR2119	UAAAAGGGAGGUGUGGAGUAG	Lj0	55824002..55824022 (-)	LIS
Cicer arietinum	miR2119	UCAAAGGGGGGUGAGGAGUAG	Ca2	22138566..22138586	LIS
Cicer reticulatum	miR2119	UCAAAGGGGGGUGAGGAGUAG	CM010872	22687608..22687628	NCBI
Cicer echinospermum	miR2119	UCAAAGGGGG-UGAGGAGUAAA	PGTU01016578	14495..14516 (-)	NCBI
Medicago truncatula	miR2119	UCAAAGGGAGGUGUGGAGUAG	chr5	19180857..19180877 (-)	Phytozome
Trifolium pratense	miR2119a	UCAAAGGGAGGUGUGGAGUAG	Tp57577_LG2	8753814..8753834	Phytozome
	miR2119b	UCAAAGGGAGGUGUGGAGUAG	Tp57577_LG2	18586895..18586915	
Trifolium subterraneum	miR2119	UCAAAGGGAGGUGUGGAGUAG	DF973777	105429..105449	NCBI
Pisum sativum	miR2119	UCAAAGGGAGGUGUGGAGUAG	PUCA013739517	14784..14804	NCBI
Lens culinaris	miR2119	UCAAAGGGAGGUGUGGAGUAG	LcChr5	55469494..55469514 (-)	KnowPulse
Arachis duranensis	miR2119	UAAAAGUGAGGUGUAGAGUAA	Aradu.A05	99398826.. 99398846	LIS
Arachis ipaensis	miR2119	UAAAAGUGAGGUGUAGAGUAA	Araip.B05	125440050..125440070 (-)	LIS
Arachis hypogaea	miR2119a	UAAAAGUGAGGUGUAGAGUAA	Arahy.05	105338161..105338181	LIS
	miR2119b	UAAAAGUGAGGUGUAGAGUAA	Arahy.15	135468696..135468716 (-)	
Arachis monticola	miR2119a	UAAAAGUGAGGUGUAGAGUAA	CM009777	116950280..116950300 (-)	NCBI
	miR2119b	UAAAAGUGAGGUGUAGAGUAA	CM009774	7666305..7666325 (-)	NCBI
Nissolia schottii	miR2119	UCAAAGAGAGGUGUAGAGUAA	QANU01002159	196694..196714	

Note.—The table shows the name of the species, miR2119 isoforms and their sequences, the fragment and the position where this sequence is located, and the information source (NCBI, Phytozome, Legumes information System [LIS] or KnowPulse). For each sequence, the position in gray highlights the base change with respect to the sequence of *P. vulgaris.* In mapping, EST, Chr: Chromosome, contig, scaffold, or identifier number indicate assembled sequences or fragments of the genome. In position, (-) indicates the sequence is located in the opposite strand. The version of each database used can be found in the Materials and Methods.

two species encode an additional copy of miR2119. Also, within this clade, *N. schottii* presents a miR2119 sequence differing in a single position (2C) from that of *Arachis*. Remarkably, we could not identify miR2119 in the genomes of *L. angustifolius* and *L. albus* (Genistoids clade), nor in species representative of the subfamilies Caesalpinioideae (*M. pudica* and *F. albida*) and Cercidoideae (*C. canadensis*). The expression of miR2119 as a small RNA has been reported for several Legume species in the Milletioids and the Hologalegina, including *G. max* (Yan et al. 2015; Wang et al. 2019); *G. soja* (Zeng et al. 2012); *Vigna mungo* (Paul et al. 2014); *V. unguiculata* (Barrera-Figueroa et al. 2011); *P. vulgaris* (Pelaez et al. 2012); *M. truncatula* (Jagadeeswaran et al. 2009; Lelandais-Briere et al. 2009); *M. sativa* (Shu et al. 2016); *Caragana intermedia* (Zhu et al. 2013); *C. arietinum* (Garg et al. 2019). In the *Arachis* genus (Dalbergioids), no annotation of mature miR2119 has been reported. For *A. hypogaea*, we explored two small RNAseq data sets and identified the expression of mature miR2119 as a small RNA through sequence analysis of the published raw data (Chi et al. 2011; Chen et al. 2019). This finding is in agreement with the sequence that we identified as encoded in the genome. Next, we analyzed the expression of miR2119 in *Lupinus* (Genistoids), where our genomic sequence analysis suggests it is absent. For *Lupinus luteus*, the expression of miR398 was documented before (Glazinska et al. 2019), but we could not find evidence of miR2119-related sRNAs in this data set, supporting the idea that miR2119 is absent in the Genistoids. Therefore, these data suggest that miR2119 is a legume-specific miRNA only found in some clades (Millettioids, Hologalegina, and Dalbergioids) within the subfamily of the Papilionoideae; notably, this miRNA is not found in the Genistoids or in more distantly related subfamilies (Caesalpinioideae and Cercidoideae).

Next, we extended the analysis of the miR2119 precursors that we found in the genomic sequences. We performed a T-coffee sequence alignment of all the precursors for miR2119 identified (sequences in table 2, supplementary fig. S1A,

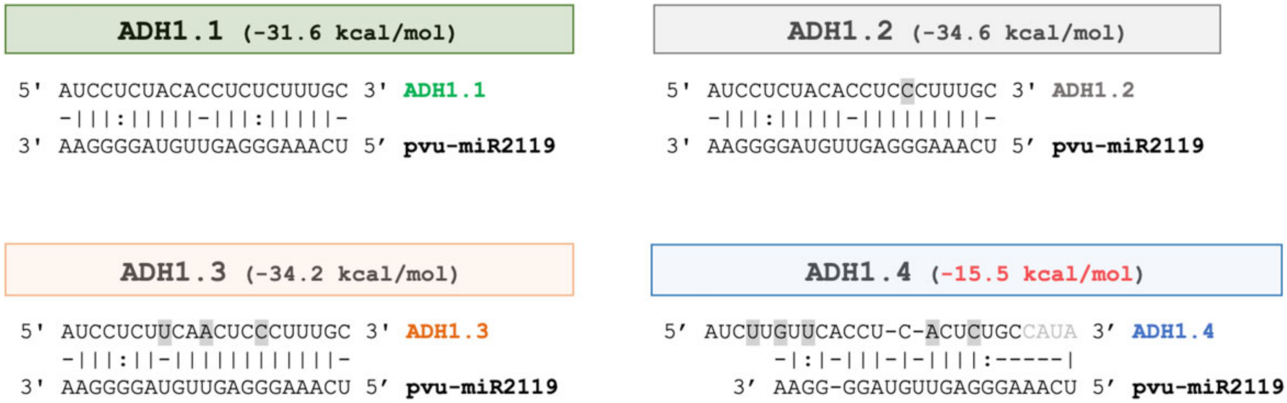

ADH1.1 (-31.6 kcal/mol)
5' AUCCUCUACACCUCUCUUUGC 3' ADH1.1
-|||:|||||-|||:|||||-
3' AAGGGGAUGUUGAGGGAAACU 5' pvu-miR2119
ADH1.2 (-34.6 kcal/mol)
5' AUCCUCUACACCUCCCUUUGC 3' ADH1.2
-|||:|||||-|||||||||-
3' AAGGGGAUGUUGAGGGAAACU 5' pvu-miR2119
ADH1.3 (-34.2 kcal/mol)
5' AUCCUCUUCAACUCCCUUUGC 3' ADH1.3
-|||:||-||||||||||||-
3' AAGGGGAUGUUGAGGGAAACU 5' pvu-miR2119
ADH1.4 (-15.5 kcal/mol)
5' AUCUUGUUCACCU-C-ACUCUGCCAUA 3' ADH1.4
-|:|-|||-|-||||:-----|
3' AAGG-GGAUGUUGAGGGAAACU 5' pvu-miR2119

Fig. 2.—miR2119 recognition site in ADH1 transcripts of *P. vulgaris*. The miR2119 binding site was identified in each of the *P. vulgaris ADH1* genes: *ADH1.1* (Phvul.009G134700), *ADH1.2* (Phvul.001G064000), *ADH1.3* (Phvul.001G067300), and *ADH1.4* (Phvul.009G149500), and the thermodynamic stability of base-pairing interaction (ΔG between ADH1: miR2119 calculated using the RNAhybrid program) is shown. Nucleotides represented in gray indicate changes based on the sequence of *ADH1.1*. Base-pairing is represented by "|," wobble pairing indicated with ":," and mismatches indicated by "-." *ADH1* gene colors represent individual members of the family, used in subsequent sections.

Supplementary Material online). This analysis revealed that, in addition to sequence conservation expected for the miRNA: miRNA* segment, a second region corresponding to the "lower stem" (the region located below the miRNA in the stem-loop structure) also revealed conserved segments. This observation is consistent with a model where the miR2119 precursor is processed in a base-to-loop manner as observed for other miRNAs as described before (Chorostecki et al. 2017). As expected, a similar analysis of miR398 precursors in the legumes (table 1) revealed a similar pattern of processing (supplementary fig. S1B, Supplementary Material online).

As described above, by analyzing the sequences of miR398 and miR2119 present in the genomes of the Papilionoideae subfamily, we found two kinds of loci encoding for miR398. In the Millettioids and Hologalegina clades, *MIR*398a is always linked to *MIR*2119. In those species that have an additional copy of *MIR*2119, such as *V. unguiculata, G. max*, and *T. pratense*, it was always associated to a *MIR*398a isoform. In contrast, when we analyzed *MIR*398a and *MIR*2119 genes in the Dalbergioids clade (*A. duranensis*, *A. ipaensis*, *A. hypogaea*, *A. monticola*, and *N. schottii*), we found that these two miRNA genes are located in separate genomic regions. These results indicate that in the Dalbergioids clade there are two loci, one encoding for *MIR*398a and another independent locus encoding for *MIR*2119 (summarized in fig. 1).

ADH1 Gene Duplication Events in the Papilionoideae Subfamily

In our previous work, the best prediction of the target mRNA for miR2119 in *P. vulgaris* was the transcript encoding for ADH1. In addition, the *ADH1* transcript was also the best candidate target for miR2119 in *P. acutifolius*, *G. max*, *M. truncatula*, *A. hypogaea*, and *L. japonicus* (De la Rosa et al. 2019). However, legumes have more than one copy of the *ADH1* gene, probably due to gene duplication events. In the *P. vulgaris* genome, there are four *ADH1* genes, which we have named as *ADH1.1* through *ADH1.4*. Three of these genes *ADH1.1, ADH1.2*, and *ADH1.3* contain a base-pairing site for miR2119 with similar thermodynamic stability values (−31.6, −34.6, and −34.2 kcal/mol, respectively, fig. 2). In addition, *ADH1.1* and *ADH1.2* were experimentally validated as miR2119 target mRNAs in *P. vulgaris* (De la Rosa et al. 2019), and related transcripts in *M. truncatula* and *G. max* (Devers et al. 2011; Shamimuzzaman and Vodkin 2012). In contrast, *P. vulgaris ADH1.4* was ruled out as a target mRNA because of the low thermodynamic stability of base-pairing to miR2119 (−15.5 kcal/mol, fig. 2).

To complement this analysis, we identified *ADH1* genes and traced their possible evolutionary history within the Papilionoideae subfamily. To this end, we obtained the protein sequences of annotated *ADH1* genes in the available genomes of species representing the Millettioids (*P. vulgaris*, *V. unguiculata*, *V. angularis*, *V. radiata*, and *C. cajan*), Hologalegina (IRLC: *M. truncatula*, *T. pratense*, and *C. arietinum*; Robinioids: *L. japonicus*), Dalbergioids (*A. duranensis*, *A. ipaensis*, and *A. hypogaea*) and Genistoids clades (*L. angustifolius*). The phylogenetic analysis of ADH1 was carried out using 67 protein sequences, including five from species outside the legumes (*A. thaliana, Prunus persicum*, *Solanum lycopersicum*, *Cucumis sativus*, and *Vitis vinifera*), and we used the ADH2 protein sequence (At5g43940.1 from *A. thaliana*) as an outgroup to root the phylogenetic tree. Based on this analysis, we defined four different clades in the Papilionoideae subfamily, each containing one of the *P. vulgaris ADH1* genes. We named these clades based on the *P. vulgaris* genes, as described in figure 3.

The ADH1.4 clade includes unique sequences from species in the Millettioids, Genistoids, and Dalbergioids clades (fig. 3,

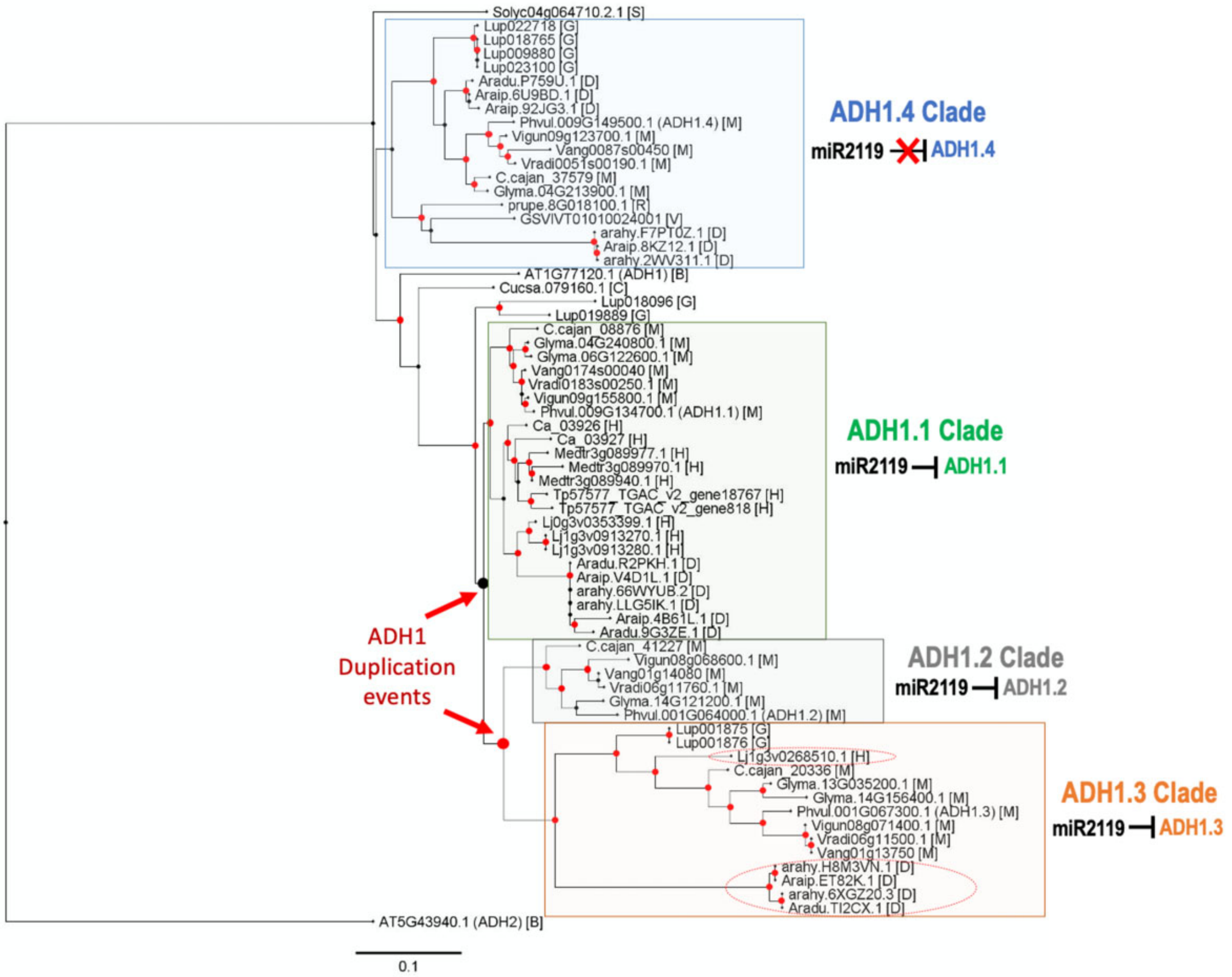


FIG. 3.—Phylogenetic analysis of ADH1 in the Papilionoideae subfamily. The phylogenetic tree was obtained based on 67 ADH1 protein sequences, which were aligned with the program MUSCLE. Afterwards, we used the ProtTest program and the phylogeny was rebuilt with the PhyML program through the *maximum likelihood* method (ML). The phylogenetic tree was visualized with the FigTree program. The sh-like values obtained for each node of the tree are represented by red dots when higher than 0.7. The black and red circles marked with arrows indicate proposed ADH1 duplication events. The clades of ADH1.1, ADH1.2, ADH1.3, and ADH1.4 are marked with a green, gray, orange, and blue rectangle, respectively. In addition, we included five ADH1 sequences of species outside the legume group including *A. thaliana* (AT1G77120.1), *P. persica* (Prupe.8G018100.1), *S. lycopersicum* (SOLYC04G064710.2.1), *C. sativus* (Cucsa.079160.1), and *V. vinifera* (GSVIVT01010024001), as well as the *A. thaliana* ADH2 protein sequence (AT5G43940.1) used as an external group for rooting of the phylogenetic tree. A red discontinuous oval shows ADH1.3 sequences that exhibit limited base-pairing with miR2119 (see text for details). Letters within brackets indicate species families as follows: Solanaceae [S], Rosaceae [R], Cucurbitaceae [C], Vitaceae [V], and Brassicaceae [B]; as well as clades: Genistoids [G], Dalbergioids [D], Hologalegina [H], and Millettioids [M]. The scale bar provides the number of substitutions per site.

blue rectangle). Other sequences that are grouped within this clade also include the nonlegumes *Prunus persica* (peach) and *V. vinifera* (grape). It is important to note that all sequences in this clade have a predicted weak base-pairing interaction with miR2119 (≥−22.6 kcal/mol), so they cannot be confidently predicted as target mRNAs for miR2119 (supplementary fig. S2, Supplementary Material online). Given the phylogenetic position of this clade, we suggest that *ADH1.4* was the first clade to diverge within the Papilionoideae subfamily whereas other clades diverged later through consecutive duplication events. For instance, within the sister group to the *ADH1.4* clade, a duplication event gave rise to the *ADH1.1* clade and the common ancestor of the *ADH1.2* and *ADH1.3* clades (see black node and upward red arrow in fig. 3); then, a subsequent duplication event led to the divergence between the *ADH1.2* and *ADH1.3* clades (see red node and downward red arrow fig. 3) during the evolution of the Papilionoideae subfamily. In the ADH1.1 clade, we identified the largest number of ADH1 sequences belonging to the Millettioids, Hologalegina, and Dalbergioids clades (fig. 3, green rectangle). For all *ADH1.1* nucleotide sequences, we observed that base-pairing to miR2119 is conserved and energetically favorable (supplementary figs. S3–S5, Supplementary Material online, Millettioids, Hologalegina, and Dalbergioids clades, respectively). Remarkably, the *ADH1.2* clade contains sequences exclusively from the Millettioids (*P. vulgaris*, *V. unguiculata*, *V. angularis*, *V. radiata*, and *C. cajan*), and all maintain a base-pairing site for miR2119 (supplementary fig. S6, Supplementary Material online), suggesting that the *ADH1.2* group emerged late in legume evolution, as it is only found in the Millettioids clade, and from an ancestor already under miR2119 regulation. In contrast, in the *ADH1.3* clade, there are sequences of the Millettioids, Hologalegina, and Dalbergioids clades, and the presence of the binding site for miR2119 is not uniform. In the Millettioids clade, each species maintains the miR2119 binding site in ADH1.3 (left panel on supplementary fig. S7, Supplementary Material online). However, the sequences from *L. japonicus* (Hologalegina) and those from *A. duranensis*, *A. ipaensis*, and *A. hypogaea* (Dalbergioids) present certain substitutions that decrease the thermodynamic stability of base-pairing to miR2119 (≥ -24.1 kcal/mol), suggesting the loss of miRNA regulation in these particular genes (right panel in supplementary fig. S7, Supplementary Material online). Finally, there are two *ADH1.3* genes in *L. angustifolius* (Lup001875 and Lup001876, Genistoids). Surprisingly, these sequences retain the binding site for miR2119 (supplementary fig. S8, Supplementary Material online), even though we could not identify miR2119 in this species. However, at this point, we cannot discard the possibility that these *ADH1.3* mRNAs could be regulated by an as-yet-unidentified miR2119 in *L. angustifolius*. The possible duplication events described thus far were evaluated using the NOTUNG program (version 2.9.1.5), which employs a parsimony criterion to infer gene transfers, duplications, and losses within gene families. The results from this analysis (see supplementary fig. S9, Supplementary Material online) not only corroborated the major duplications events shown in figure 3, but also suggested many other duplications (39 events in total) and quite a few gene losses (115 events) within this gene family.

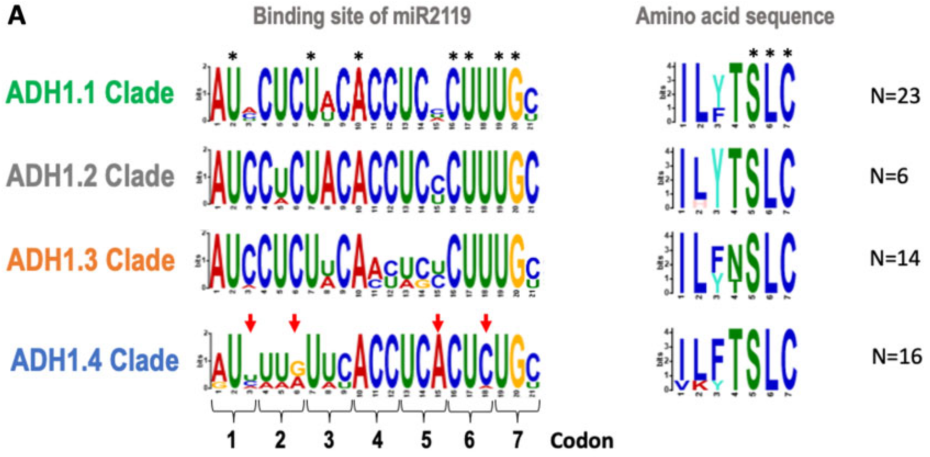
We can summarize our analysis of the presence or absence of *ADH1* genes to understand the events that lead to their current organization in the Papilionoideae subfamily. In the early branching Genistoids clade, there are *ADH1* genes (Lup018096 and Lup019889 in *L. angustifolius*) that we infer gave rise to ADH1.1, as well as to the ancestor of ADH1.2 and ADH1.3 (fig. 3 and supplementary fig. S10, Supplementary Material online). Accordingly, the Dalbergioids clade contains the sequences of ADH1.1, ADH1.3, and ADH1.4. In the Hologalegina clade, *L. japonicus* of the Robinioide subclade, presents ADH1.1 and ADH1.3, with the possible loss of ADH1.4, whereas the IRLC subclade only contains multiple copies of ADH1.1, suggesting the loss of ADH1.3 and ADH1.4. Finally, the Millettioids clade contains sequences encoding ADH1.1, ADH1.3, and ADH1.4, and interestingly, we detected *ADH1.2*, a gene unique to this clade, which suggests its late emergence (supplementary fig. S10, Supplementary Material online). Altogether, each species of the Millettioids, Hologalegina, and Dalbergioids shares at least one copy of *ADH1.1* regulated by miR2119.

The Recognition Site for miR2119 Is Conserved in *ADH1* Genes Independently of Amino Acid Sequence Requirements

In plants, the binding site for an miRNA can be located throughout the target transcript, in the 5′UTR, in the coding region or in the 3′UTR (Brodersen et al. 2008). The miR2119 binding site in *ADH1* transcripts is located in the coding region; thus, its sequence conservation may be determined by the selection pressure operating at the nucleotide level to maintain the recognition by miR2119, as well as by the amino acid identity in the protein sequence. To dissect the contribution of these two factors, we first determined the consensus for the nucleotide and amino acid sequences corresponding to the miR2119 binding site for each of the ADH1.1, ADH1.2, ADH1.3, and ADH1.4 clades.

The consensus sequences obtained show a high similarity and conservation between the ADH1.1 and ADH1.2 clades at the nucleotide level (fig. 4*A*), notably both target mRNAs were validated experimentally in *P. vulgaris* before (De la Rosa et al. 2019). The consensus sequence of ADH1.3 shows considerable variation at positions 11–15, and these changes cause an extended mismatched region in the ADH1.3:miR2119 interaction (fig. 4*A* and right panel in supplementary fig. S7, Supplementary Material online). Remarkably, the nucleotide consensus sequence for ADH1.4 has a larger number of variations with respect to the other ADH1 clades, as it shows poor conservation in positions 3–6 and 8–9, and contains at least four positions completely different from ADH1.1, ADH1.2, and ADH1.3 (pos. 4, 6, 15, and 18, fig. 4*A*, marked with red arrows). Taken together, these results indicate that miR2119 has perfect binding sites in ADH1.1 and ADH1.2, a slightly degenerate site in ADH1.3, but a nonfunctional binding site in ADH1.4 (representative miR2119 sites as those present in *P. vulgaris ADH1* genes are shown in fig. 4*B*).

Despite the differences at the nucleotide level shown in the sequence corresponding to the miR2119 binding site, the corresponding amino acid consensus sequences in the four ADH1 clades show high degree of similarity to each other (right panel in fig. 4*A*). The 21-nt binding site matches the +1 open reading frame for protein translation, encoding for seven amino acid residues located in the catalytic domain of the protein. The amino acids at positions 5–7 (Ser, Leu, and Cys, respectively) are highly conserved in angiosperms with



B

```
5' AUCCUCUUCAACUCCCUUUGC 3' ADH1.3 (-34.2 kcal/mol)
5' AUCCUCUACACCUCCCUUUGC 3' ADH1.2 (-34.6 kcal/mol)
5' AUCCUCUACACCUCUCUUUGC 3' ADH1.1 (-31.6 kcal/mol)
   -|||:|||||-|||:|||||-
3' AAGGGGAUGUUGAGGGAAACU 5' pvu-miR2119
   -||::-|-||-|||-||-||-
5' AUCUUGUUCACCUCACUCUGC 3' ADH1.4
```

Fig. 4.—miR2119 binding sites in the four ADH1 clades reflect their corresponding selection factors. (*A*) Consensus sequence sites for miR2119 in ADH1.1, ADH1.2, ADH1.3, and ADH1.4. Left panel shows the binding site in nucleotides and right panel displays consensus site in the corresponding amino acid residues. Horizontal key brackets numbered 1–7 indicate the codon positions for amino acids in the sequence of ADH1. "N" indicates the number of sequences used to obtain each consensus using the MEME suite. Asterisks indicate invariable positions and red arrows show positions in ADH1.4 that affect base-pairing with miR2119. (*B*) Base-pairing interaction of each copy of ADH1 in *P. vulgaris* with miR2119, with gray boxes showing base changes with respect to the ADH1.1 sequence. Nucleotides in red indicate a base change that causes a mismatch between ADH1 and miR2119. Base-pairing is represented by "|," wobble pairing indicated with ":," and mismatch indicated by "-." Base-pairing of miR2119 to ADH1.1, ADH1.2, and ADH1.3 is very similar, and thus it is represented only once by showing the interaction between miR2119 and ADH1.1.

the cysteine residue being important for binding of a zinc ion, used as a cofactor by this enzyme (Strommer 2011). However, the nucleotide consensus of ADH1.4 shows synonymous substitutions in the third position of codons 5 and 6 that are incompatible with the regulation by miR2119 but maintain the identity of the encoded amino acid residues. By contrast, these positions remain unchanged in ADH1.1, ADH1.2, and ADH1.3, strongly suggesting an additional selection pressure at the nucleotide level in these genes to maintain the regulation by miR2119 (fig. 4*A*). Thus, these results show that the sequence of the miR2119 binding site is under selective pressure by at least two independent factors, first at the nucleotide level to retain regulation by miR2119 and second, to preserve the amino acid sequence necessary for enzyme activity.

Discussion

There are different models to explain the varied origins of new miRNAs in plants. One such model entails the duplication of the gene encoding the future target mRNA to generate a partial inverted repeat. The transcript originating from this new locus then adopts a perfectly complementary secondary structure, which is substrate of double-stranded RNA endonucleases of the Dicer-like family such as DCL3 or DCL4 to generate multiple small RNAs (siRNA, *small interfering RNA*). In turn, these siRNAs regulate the expression of the transcript of origin, as well as those of homologous genes. Over time, the novel partial inverted repeat gene accumulates mutations that allow the double-stranded RNA to be recognized as an miRNA precursor and to be processed by DCL1, giving rise to a new miRNA (Allen et al. 2004; Cui et al. 2017; Baldrich et al. 2018).

A handful of examples has emerged to provide support to the model of partial tandem gene duplication encoding for a target mRNA as a generator for new miRNAs. One such case involves the large family of Nucleotide-binding site Leucine-rich repeat (NBS-LRR) receptors associated to pathogen defense responses and widely distributed in both monocotyledonous and dicotyledonous plants. At least eight different miRNA families have been described as regulators of the

NBS-LRR genes, where a common attribute among them is the conservation of the sequence that serves as binding site on target mRNAs, allowing the regulation of multiple-related genes using a single miRNA (Fei et al. 2016). For example, members of the miR482/2118 family recognize the site encoding for the conserved P-Loop motif present in the NBS-LRR (Shivaprasad et al. 2012). Recently, it was observed that high duplication frequency in the different families of NBS-LRR genes was associated with the emergence of a novel miRNA. This was supported by the extensive similarity observed between the miRNA precursor sequence and the sequence of its target NBS-LRR genes (Zhang et al. 2016). Other lineage-specific miRNAs with similar characteristics include *MIR*472, *MIR*825, and *MIR*1885 in Brassicaceae; *MIR*1510 and *MIR*2089 in Fabaceae; *MIR*6025 in Solanaceae, *MIR*5163 and *MIR*9863 in Poaceae (Zhang et al. 2016), suggesting that similar duplication events have occurred in different plant families. To address this possibility for *MIR*2119, we explored the sequences of the miRNA precursors and their similarity to *ADH1* genes. The *MIR*2119 precursor sequences obtained for Millettioids, Hologalegina, and Dalbergioids were separated into shorter regions considering their conservation. Each sequence was then used as a query to search for limited similarities with *ADH1* genes or any other genomic regions. Despite adjusting some parameters to allow for nucleotide mismatches, our sequence comparison did not reveal any clear similarities between the precursor of miR2119 and the *ADH1* genes in several Papilionoideae analyzed, yet this could be due to accumulated mutations in the precursor during the long-elapsed time since its origin.

Independently of the mechanism that gave rise to the *MIR*2119 gene within the Papilionoideae subfamily, we propose it originated in the common ancestor of the Dalbergioids and Hologalegina-Millettioids clades, ca. 55–56 Ma, according to the commonly accepted evolutionary history of the Papilionoideae (Lavin et al. 2005). Within the Dalbergioids, an *MIR*2119 locus is present in species belonging to the genera *Arachis* and *Nissolia* as an independent transcription unit (fig. 1). In contrast, *MIR*2119 was not identified in *L. angustifolius* and *L. albus*, representatives of the Genistoids clade, and neither in earlier diverging species, such as *M. pudica*, *F. albida*, and *C. canadensis*, which belong to the Caesalpinioideae and Cercidoideae subfamilies, respectively. However, we cannot rule out that *L. angustifolius*, *L. albus*, *M. pudica*, *F. albida*, and *C. canadensis* have suffered the loss of miR2119 or that sequencing errors in the annotation of these genomes precluded its identification. Alternatively, the sequence of miR2119 in these species could be so different from the one detected here, that it prevented its recognition. The future availability of genome sequences and more small RNA-sequencing data for related species will help to clarify this issue.

Polycistronic miRNA precursors in plants can have different evolutionary origins. In the case of *MIR*395, tandem homologous miRNAs are present in the same transcript in different species, which could have originated from multiple duplication events. This arrangement results in larger miRNA accumulation and consequently in a larger dose effect on the repression of its target mRNA(s) (Guddeti et al. 2005; Nozawa et al. 2012). A different scenario has been described for polycistronic precursors containing nonhomologous miRNAs. It has been proposed that these polycistronic precursors originated from a partial gene duplication event, where the duplicated inverted fragment would be large enough to generate two new miRNA precursors with different sequence. As the new miRNAs originated from a single source, they end up regulating transcripts from the same or similar gene family (Merchan et al. 2009).

During the study of *MIR*398 and *MIR*2119, we identified the presence of a dicistronic precursor gene in the Hologalegina and Millettioids clades. The acquisition of a *MIR*398–*MIR*2119 gene probably occurred in their common ancestor ca. ∼50–55 Ma (Lavin et al. 2005), after the separation from the Dalbergioids clade, which already contained an independent *MIR*2119 locus (in *Arachis* and *Nissolia* genera). This event probably originated through a process of genomic rearrangement that caused the fusion of two genes initially separated and that allowed the cotranscription of both miRNAs, showing a new mechanism for the generation of polycistronic miRNA genes. We speculate that this rearrangement created new opportunities for the spatial and temporal coordination of the expression of their target mRNAs, CSD1 and ADH1; likely contributing to a better coupling of the corresponding enzymatic activities according to the adaptive metabolic needs of the legume species involved.

In our study, we confirmed that *MIR*2119 is only found in species of the Millettioids, Hologalegina, and Dalbergioids clades within the Papilionoideae subfamily. Given the *MIR*2119 species distribution, in conjunction with the phylogenetic analysis of the *ADH1* genes, we propose that the emergence of *MIR*2119 probably occurred during the duplication processes involving its future target genes (fig. 5). In our model, an ancestral *ADH1* gene lacking an miR2119 binding site gave rise to *ADH1.1* by gene duplication. Because *ADH1.1* is shared in species containing miR2119, it is possible that the miRNA emerged through the doubling model of "the gene in tandem" (opposite orientation) by partial duplication of *ADH1.1*. Transcription of this new gene generated a perfectly complementary double-stranded RNA, capable of DCL processing to generate siRNAs targeting *ADH1.1* transcripts. After its emergence, the novel gene accumulated point mutations leading to the production of a functional precursor encoding miR2119. In consequence, paralogous genes emerging from *ADH1.1* would then be subjected to miR2119 regulation (fig. 5).

Finally, we observed that the complex combination of *ADH1* genes in the different Papilionoideae clades correlates with the presence of *MIR*2119 (supplementary fig. S10,

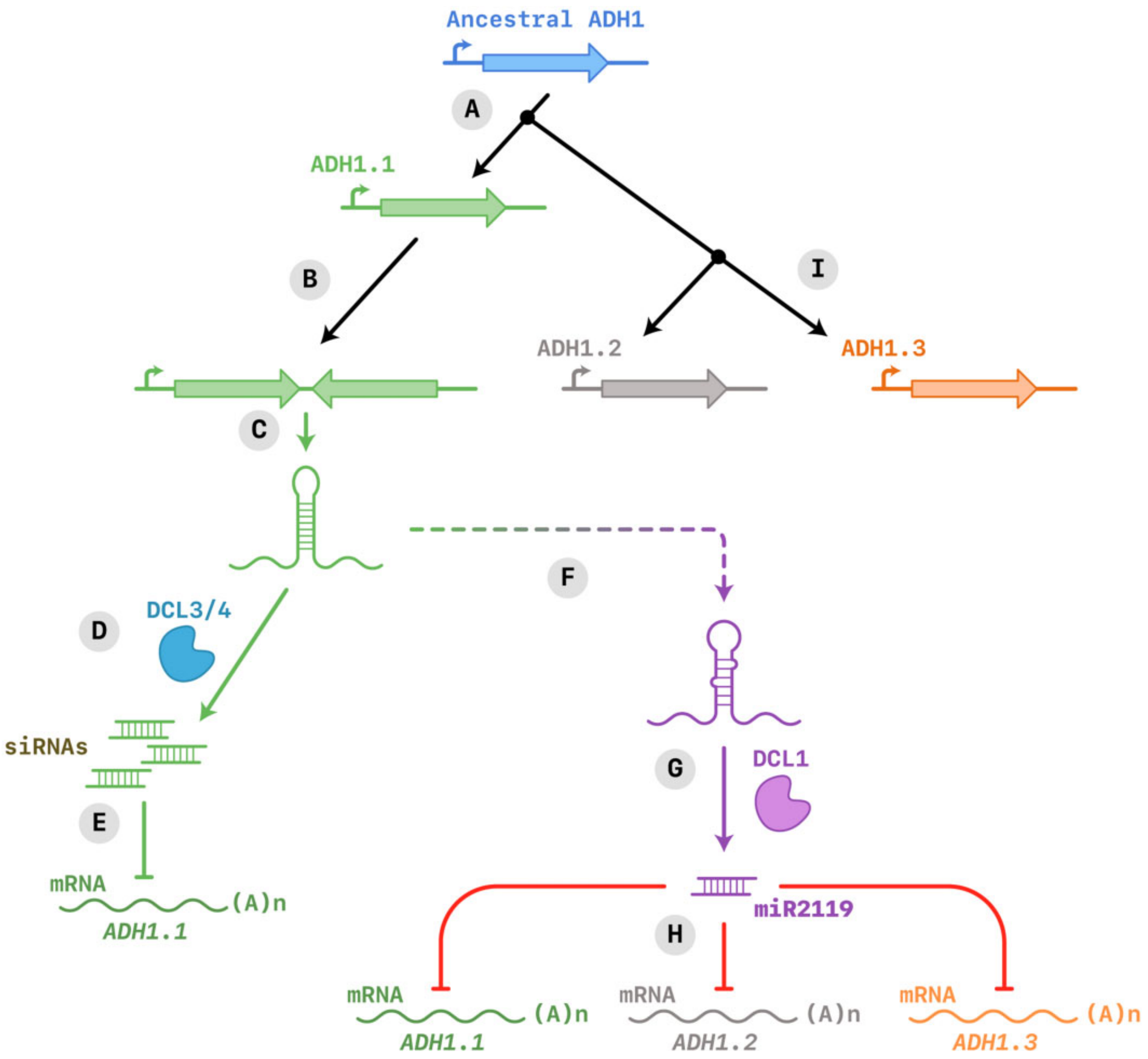


Fig. 5.—Possible scenario for the origin and evolution of miR2119 and its regulatory target genes. (*A*) We suggest that a pre-existing copy of an ancestral *ADH1* gene diverged to give rise to a second locus, here shown as ADH1.1 by a gene duplication event. (*B*) In turn, we suggest that a partial duplication of this gene generated an inverted repeat in a convergent direction. (*C*) Transcription of the inverted gene led to formation of a fully complementary double-stranded RNA. (*D*) In turn, double-stranded RNA processed by DCL3 or DCL4 generated multiple small interfering RNAs (siRNAs) that inhibited the expression of the transcript of the gene of origin (*E*). (*F*) The accumulation of mutations in the siRNA-generating locus caused imperfect complementarity in the double-stranded RNA and led to formation of a new miRNA precursor (pre-miR2119). (*G*) Recognition and processing of the miR2119 precursor by DCL1. (*H*) Generation of mature miR2119 that can regulate the transcript of the gene of origin (*ADH1.1*) or related mRNAs (such as ADH1.2 or ADH1.3) that originated from other gene duplication events (*I*).

Supplementary Material online). This fact suggests that these two elements could be closely linked. As discussed above, it remains to be determined if *ADH1* gene rearrangements were responsible for *MIR*2119 emergence in the Papilinioideae. At a different level, miR2119 regulation constrains the abundance of *ADH1* gene transcripts containing miRNA binding sites, but not of other transcripts, such as *ADH1.4*. In this way, the presence of miR2119 in a given genome may affect the number and kind of *ADH1* genes present, suggesting another layer of complexity to the evolutionary history of the *ADH1*-*MIR*2119 module.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

C.D.l.R. first conceived this work. C.D.l.R. and L.L. retrieved sequence data and performed phylogenetic analyses. C.D.l.R., L.L., S.C., A.C.R., and J.L.R. discussed results and planned analyses. C.D.l.R. and J.L.R. wrote the first manuscript draft with input from all authors. We are grateful to Blake C.

Meyers for his comments on this manuscript. We thank Alfredo Hernández, Romualdo Zayas, and Víctor del Moral at CCG-UNAM for technical support and Pascual Redondo for help with graphic design. This work was supported by a grant from DIrección General de Asuntos del Personal Académico-Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (DGAPA-PAPIIT), Universidad Nacional Autónoma de México (UNAM) project number IN202918.

Data Availability

The data underlying this article are available in the article and in its online supplementary material.

Literature Cited

Allen E, et al. 2004. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. Nat Genet. 36(12):1282–1290.

Arenas-Huertero C, et al. 2009. Conserved and novel miRNAs in the legume *Phaseolus vulgaris* in response to stress. Plant Mol Biol. 70(4):385–401.

Axtell MJ. 2013. Classification and comparison of small RNAs from plants. Annu Rev Plant Biol. 64(1):137–159.

Azani N, et al. 2017. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny—the Legume Phylogeny Working Group (LPWG). Taxon 66(1):44–77.

Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37(Web Server):W202–W208.

Baldrich P, Beric A, Meyers BC. 2018. Despacito: the slow evolutionary changes in plant microRNAs. Curr Opin Plant Biol. 42:16–22.

Barrera-Figueroa BE, et al. 2011. Identification and comparative analysis of drought-associated microRNAs in two cowpea genotypes. BMC Plant Biol. 11(1):127.

Brodersen P, et al. 2008. Widespread translational inhibition by plant miRNAs and siRNAs. Science 320(5880):1185–1190.

Bui LT, et al. 2019. Conservation of ethanol fermentation and its regulation in land plants. J Exp Bot. 70(6):1815–1827.

Chen H, et al. 2019. Integrated microRNA and transcriptome profiling reveals a miRNA-mediated regulatory network of embryo abortion under calcium deficiency in peanut (*Arachis hypogaea* L.). BMC Genomics 20(1):392.

Chi X, et al. 2011. Identification and characterization of microRNAs from peanut (*Arachis hypogaea* L.) by high-throughput sequencing. PLoS One 6(11):e27530.

Chorostecki U, et al. 2017. Evolutionary footprints reveal insights into plant microRNA biogenesis. Plant Cell 29(6):1248–1261.

Cui J, You C, Chen X. 2017. The evolution of microRNAs in plants. Curr Opin Plant Biol. 35:61–67.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27(8):1164–1165.

Dash S, et al. 2016. Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family. Nucleic Acids Res. 44(D1):D1181–D1188.

De la Rosa C, Covarrubias AA, Reyes JL. 2019. A dicistronic precursor encoding miR398 and the legume-specific miR2119 coregulates CSD1 and ADH1 mRNAs in response to water deficit. Plant Cell Environ. 42(1):133–144.

*Devers EA, Branscheid A, May P, Krajinski F. 2011. Stars and symbiosis: microRNA- and microRNA-mediated transcript cleavage involved in arbuscular mycorrhizal symbiosis. Plant Physiol. 156(4):1990–2010.

Di Tommaso P, et al. 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 39(Suppl):W13–W17.

Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. 2012. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics 13(1):162.

Doyle JJ, Luckow MA. 2003. The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol. 131(3):900–910.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Fei Q, Zhang Y, Xia R, Meyers BC. 2016. Small RNAs add zing to the Zig-Zag-Zig model of plant defenses. Mol Plant Microbe Interact. 29(3):165–169.

Garg V, et al. 2019. Integrated transcriptome, small RNA and degradome sequencing approaches provide insights into *Ascochyta* blight resistance in chickpea. Plant Biotechnol J. 17(5):914–931.

Gepts P, et al. 2005. Legumes as a model plant family. Plant Physiol. 137(4):1228–1235.

Glazinska P, Kulasek M, Glinkowski W, Wojciechowski W, Kosinski J. 2019. Integrated analysis of small RNA, transcriptome and degradome sequencing provides new insights into floral development and abscission in yellow lupine (*Lupinus luteus* L.). Int J Mol Sci. 20:5122.

Graham PH, Vance CP. 2003. Legumes: importance and constraints to greater use. Plant Physiol. 131(3):872–877.

Griesmann M, et al. 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. Science 361:eaat1743.

Guddeti S, et al. 2005. Molecular evolution of the rice miR395 gene family. Cell Res. 15(8):631–638.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 537:113–137.

Jagadeeswaran G, et al. 2009. Cloning and characterization of small RNAs from *Medicago truncatula* reveals four novel legume-specific microRNA families. New Phytol. 184(1):85–98.

Jones-Rhoades MW, Bartel DP. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell. 14(6):787–799.

Kruger J, Rehmsmeier M. 2006. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res. 34(Web Server issue):W451–W454.

Lavin M, Herendeen PS, Wojciechowski MF. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. Syst Biol. 54(4):575–594.

Lelandais-Briere C, et al. 2009. Genome-wide *Medicago truncatula* small RNA analysis revealed novel microRNAs and isoforms differentially regulated in roots and nodules. Plant Cell. 21(9):2780–2796.

Merchan F, Boualem A, Crespi M, Frugier F. 2009. Plant polycistronic precursors containing non-homologous microRNAs target transcripts encoding functionally related proteins. Genome Biol. 10(12):R136.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302(1):205–217.

Nozawa M, Miura S, Nei M. 2012. Origins and evolution of microRNA genes in plant species. Genome Biol Evol. 4(3):230–239.

Paul S, Kundu A, Pal A. 2014. Identification and expression profiling of *Vigna mungo* microRNAs from leaf small RNA transcriptome by deep sequencing. J Integr Plant Biol. 56(1):15–23.

Pelaez P, et al. 2012. Identification and characterization of microRNAs in *Phaseolus vulgaris* by high-throughput sequencing. BMC Genomics 13(1):83.

Posada D. 2009. Selection of models of DNA evolution with jModelTest. Methods Mol Biol. 537:93–112.

Shamimuzzaman M, Vodkin L. 2012. Identification of soybean seed developmental stage-specific and tissue-specific miRNA targets by degradome sequencing. BMC Genomics 13(1):310.

Shivaprasad PV, et al. 2012. A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. Plant Cell 24(3):859–874.

Shu Y, et al. 2016. Genome-wide investigation of microRNAs and their targets in response to freezing stress in *Medicago sativa* L., based on high-throughput sequencing. G3 (Bethesda) 6(3):755–765.

Sievers F, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 7:539.

Stolzer M, et al. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics 28(18):i409–i415.

Strommer J. 2011. The plant ADH gene family. Plant J. 66(1):128–142.

Sunkar R, Zhu JK. 2004. Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. Plant Cell 16(8):2001–2019.

Wang R, et al. 2019. Construction and analysis of degradome-dependent microRNA regulatory networks in soybean. BMC Genomics 20(1):534.

Xu S, Guerra D, Lee U, Vierling E. 2013. S-nitrosoglutathione reductases are low-copy number, cysteine-rich proteins in plants that control multiple developmental and defense responses in Arabidopsis. Front Plant Sci. 4:430.

Yan Z, et al. 2015. Identification of microRNAs and their mRNA targets during soybean nodule development: functional analysis of the role of miR393j-3p in soybean nodulation. New Phytol. 207(3):748–759.

Zeng QY, et al. 2012. Identification of wild soybean miRNAs and their target genes responsive to aluminum stress. BMC Plant Biol. 12(1):182.

Zhang Y, Xia R, Kuang H, Meyers BC. 2016. The diversification of plant NBS-LRR defense genes directs the evolution of microRNAs that target them. Mol Biol Evol. 33(10):2692–2705.

Zhu C, Ding Y, Liu H. 2011. MiR398 and plant stress responses. Physiol Plant. 143(1):1–9.

Zhu J, Li W, Yang W, Qi L, Han S. 2013. Identification of microRNAs in *Caragana intermedia* by high-throughput sequencing and expression analysis of 12 microRNAs and their targets under salt stress. Plant Cell Rep. 32(9):1339–1349.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31(13):3406–3415.

Associate editor: Vision Todd