ARTICLE - DISCOVERIES

Genome-wide analysis of small RNA biogenesis proteins refine the evolution of Dicer-like and Argonaute gene families in flowering plants

Sébastien Bélanger[1§], Junpeng Zhan[1§], Blake C. Meyers[1,2*]

[1] Donald Danforth Plant Science Center, St. Louis, MO, USA, 63132

[2] Division of Plant Science and Technology, University of Missouri, Columbia, MO, USA, 65211

[§] These authors contributed equally

* Corresponding author: BMeyers@danforthcenter.org

Running title: *Evolution of plant small RNA pathways*

One sentence summary: Analysis of small RNA biogenesis proteins from over 198 plant species refines the evolution of Dicer-like, Argonaute and other gene families, with insights into their expansion in flowering plants.

ABSTRACT

Several protein families play a role in the biogenesis and function of small RNAs (sRNAs) in plants. Among these are Dicer-like (DCL), RNA-directed RNA polymerase (RDR), and Argonaute (AGO) proteins, as well as accessory proteins such as double-stranded RNA-binding (DRB), SERRATE (SE), and SUPPRESSION OF SILENCING 3 (SGS3) that are co-factors of DCL and RDR proteins. Here, we present phylogenetic analyses of seven sRNA pathway protein families performed on 198 species in the Viridiplantae. We show that the eudicot DCL3 protein is orthologous to the monocot DCL5, indicating that the monocot DCL3 has diverged and specialized in heterochromatic siRNA production. We found 167 highly conserved amino acids conserved among monocot DCL3/DCL4/DCL5 sequences. Amino acids conserved between DCL3 and DCL5 were located in the Helicase and RNase domains while amino acids conserved between DCL4 and DCL5 were mostly in the dsRNA binding and PAZ domains. Absent in DCL4, we found protein regions common to DCL3/DCL5 near both extremities of the PAZ domain which can potentially increase the distance from the RNase III catalytic center and thus determine the product size. We identified multiple duplication events of AGOs that had been lost, retained or further duplicated in sub-groups, indicating that the evolution of AGOs is complex in monocots, suggesting that functional studies in model species might not be easily translated to other species. This work provides an evolutionarily coherent and standardized annotation for gene families involved in plant sRNA biogenesis.

INTRODUCTION

Numerous plant species are known to produce small non-coding RNAs (sRNAs), which are usually 20 to 24 nucleotides (nt) in length (Axtell 2013). The best characterized sRNAs in plants are microRNAs (miRNAs) and short interfering RNAs (siRNAs). Typically, miRNAs originate from single-stranded transcripts produced by RNA polymerase II (Pol II). The transcripts fold into hairpin-like structures that are cleaved by a Dicer-like protein called DICER-LIKE 1 (DCL1) to produce miRNAs, which are usually 21 or 22 nt in length (Borges and Martienssen 2015). miRNAs regulate many biological processes through post-transcriptional or translational regulation of target mRNAs or long non-coding RNA transcripts (Borges and Martienssen 2015; Axtell and Meyers 2018). Generally, 21-nt miRNAs mediate cleavage of protein-coding RNAs to trigger their degradation, while 22-nt miRNAs trigger the cleavage of protein-coding or non-coding RNAs to initiate the production of 21-nt or 24-nt, phased, small interfering RNAs (phasiRNAs) (Borges and Martienssen 2015). siRNAs can be categorized into heterochromatic siRNAs (hc-siRNAs) and phased siRNAs (phasiRNAs). hc-siRNAs are typically 24 nt in length and their precursors are produced from intergenic, repetitive, and transposon-related genomic regions by Pol IV (Axtell and Meyers 2018). hc-siRNAs recognize specific genomic regions and guide de novo deposition of DNA methylation (i.e., RNA-directed DNA methylation [RdDM]) on adjacent regions to regulate transcription (Borges and Martienssen 2015; Axtell and Meyers 2018). The second group of siRNAs, phasiRNAs, originate from protein-coding transcripts or long non-coding RNAs (lncRNAs), which are Pol II products (Fei et al. 2013; Liu et al. 2020). PhasiRNA biogenesis initiates via miRNA-directed, Argonaute (AGO)-catalyzed cleavage of a single-stranded RNA (ssRNA) precursor, which is then converted to a double-stranded RNA (dsRNA) by an RDR protein, and processed into 21-nt or 24-nt RNA duplexes by a DCL protein. PhasiRNAs can be further sub-classified as trans-acting siRNAs and reproductive phasiRNAs (derived from long non-coding loci and are enriched in male reproductive tissues) (reviewed in (Liu et al. 2020)) . Both types of phasiRNAs have been shown to mediate cleavage of target transcripts (Fei et al. 2013; Tamim et al. 2018; Jiang et al. 2020; Zhang et al. 2020); a subset of each type has also been implicated in regulating DNA methylation (Wu et al. 2012; Zhang et al. 2021). The

diversity of origins and activities of plant sRNAs are likely associated with the presence of specific biogenesis or effector machineries.

In addition to the RDR, DCL, and AGO proteins mentioned above, several other protein families are involved in small RNA biogenesis and/or activities: the double-stranded RNA-binding (DRB) and SERRATE (SE) proteins are known to form complexes with DCL proteins, SGS3 is a well-known protein partner of RDR6, while the HUA ENHANCER 1 (HEN1) protein, encoding an RNA methyltransferase, is involved in small RNA modifications (Z. Yang et al. 2006). Each of the protein families in Arabidopsis (*Arabidopsis thaliana*) have specific conserved domains (Supplemental Figure 1). Pathways governing miRNA biogenesis and activity likely emerged in the common ancestor of land plants, whereas the origin of the machinery of hc-siRNA biogenesis is less understood because the components are frequently lost in sister species of seed plants (You et al. 2017; S. Wang et al. 2021).

At least two whole genome duplication (WGD) events occurred prior to the diversification of angiosperms; one near the origin of the seed plants and one near the origin of the angiosperms (Jiao et al. 2011; Jiao et al. 2012; Schranz et al. 2012). Another occurred later near the origin of the eudicots while yet another has been suggested at the base of the Poaceae (Paterson et al. 2004; Salse et al. 2008; Schranz et al. 2012). More recent polyploidy events arose in many species such as *Nicotiana tobaccum*, *Triticum* ssp or *Avena sativa* (Dubcovsky and Dvorak 2007; Edwards et al. 2017; IWGSC 2018; Maccaferri et al. 2019; Peng et al. 2021). Genome duplications can lead to evolutionary changes that (i) silence a gene copy (pseudogenization), (ii) result in a paralog with a novel function (neofunctionalization), or (iii) generate paralogs that partition an original function (subfunctionalization) (Conant and Wolfe 2008; Hahn 2009; Pozo and Ramirez-Parra 2015). Gene duplication has driven the functional diversification of several proteins involved in the biogenesis or functional machinery of sRNAs, namely AGO, DCL and RDR proteins (and presumably the DRB proteins) that are involved in the biogenesis of phasiRNAs (Table 1). Biogenesis of hc-siRNAs requires the activity of RDR2, DCL3 and AGO4/6 (Borges and Martienssen 2015), whereas the production of phasiRNAs involves the activity of RDR6 and DCL4 (21-nt

phasiRNAs) or DCL5 (24-nt phasiRNAs) (Liu et al. 2020). Described as unique to monocots, DCL5 (initially called DCL3b) evolved from DCL3 (Patel et al. 2021). Both DCL3 and DCL5 proteins process dsRNA substrates to yield 24-nt sRNAs, but DCL3 generates a single sRNA duplex while DCL5 catalyzes a recursive cleavage in regular duplex increment (Herr et al. 2005; Liu et al. 2020). Consistent with the role of DCL4 in producing 21-nt sRNAs, the DCL4 orthologs form a distinct sister clade to DCL3/5 (Liu et al. 2020). However, interestingly, DCL4 cleaves dsRNAs in a manner similar to DCL5; that is, it also catalyzes a recursive cleavage in regular duplex increment (Liu et al. 2020). Together, these observations raise several questions. How did DCLs evolve to (i) a DCL5 protein that processes recursive cleavage, like DCL4, but apparently not DCL3, and (ii) a DCL4 protein that processes phasiRNA duplexes in 21-nt increments while DCL5 processes in 24-nt increments? Also, in the absence of DCL5, which DCL protein produces 24-nt phasiRNAs in eudicots? Another protein family vital to sRNA activity are AGO proteins. In the monocot reproductive phasiRNA pathways, an AGO1 protein presumably forms a complex with miR2118 or miR2275 to initiate the biogenesis of 21-nt and 24-nt phasiRNAs by catalyzing miRNA-triggered cleavage of precursor transcripts. Then, additional AGO proteins load the phasiRNAs; AGO5 orthologs load 21-nt phasiRNAs in rice (*Oryza sativa*) and maize (*Zea mays*) (Komiya et al. 2014; Lee et al. 2021), and the maize AGO18b loads both 21-nt and 24-nt phasiRNAs (Sun et al. 2019).

In this study, we performed phylogenetic analyses and characterized seven protein families involved in the regulation of plant sRNA biogenesis in 198 Viridiplantae species following available gene annotation nomenclatures for Arabidopsis, rice and maize. We describe the phylogenetic relationships of these proteins in the plant kingdom. We show that the eudicot DCL3 is sister to the monocot DCL5 clade and that the monocot *DCL3* gene resulted from a duplication and divergence in the monocot lineage, yielding subfunctionalized and specialized proteins not observed in eudicots. We examined the conservation and divergence of protein sequence and biochemical structure between DCL3/4/5 proteins found in monocots and observed 167 highly conserved amino acids. Among those, 46 amino acids conserved between DCL3 and DCL5 were located in the Helicase and RNase domains while the amino acids conserved between DCL4 and

DCL5 were mostly in the dsRNA and PAZ domains. The protein regions that are present in both DCL3 and DCL5 but absent in DCL4 were found near both extremities of the PAZ domain, and can potentially increase the distance from the RNase III catalytic center and thus the size of sRNA products. In addition, we found evidence that *AGO* genes were lost, retained, or further duplicated in sub-groups of monocots, indicating complex patterns of AGO evolution in the lineage. This work results in an evolutionarily coherent and standardized annotation for genes encoding several sRNA biogenesis/effector protein families, and provides insights into the evolution of several sRNA pathways, including the reproductive phasiRNA pathways.

RESULTS AND DISCUSSION

Annotation of sRNA pathway genes across species in the Viridiplantae kingdom

To better understand the evolution of sRNA pathways in plants, we performed a phylogenetic analysis on several relevant protein families across a diverse set of species. We analyzed 219 species (Supplemental Table 1) including 11 Chlorophyta and 187 Embryophyta species, and 21 outgroup species (including five Rhodophyta, one Oomycota, one Amoebozoa, four Fungi and 10 Metazoa). The Embryophyta species included 181 Tracheophyta species comprising one Lycopodiopsida (lycophyte), two Polypodiopsida (ferns), eight Acrogymnospermae (gymnosperms) and 170 Mesangiospermae (angiosperms; including two basal angiosperms, one Magnoliids, 48 Liliopsida (monocots) and 119 Pentapetalae (eudicots)) (Supplemental Figure 2). The proteins we examined included DICER-LIKE (DCL), DOUBLE-STRANDED RNA-BINDING (DRB), SERRATE (SE), HUA ENHANCER 1 (HEN1), RNA-DIRECTED RNA POLYMERASE (RDR), SUPPRESSION OF SILENCING 3 (SGS3) and ARGONAUTE (AGO) families (Supplemental Figure 1). Together, a total of 2,895 AGO, 1,073 DCL, 1,886 DRB, 276 HEN1, 1,349 RDR, 469 SE and 465 SGS3 proteins in the 198 plant species were annotated and curated for the presence of functional domains.

Molecular evolution of RDR proteins

RDR proteins are responsible for converting ssRNA to dsRNA, which is a key step in siRNA biogenesis (Borges and Martienssen 2015). Thus, we performed a phylogenetic analysis of the RDR protein and explored patterns of *RDR* duplication and loss among plant genomes. A maximum likelihood tree of 1,375 RDR proteins (including 26 non-plant RDRs) was constructed and rooted with fungal and algal species as the outgroup (Figure 1; Supplemental data 1; Supplemental table 2). We detected RDR-like proteins in all fungi (4 of 4 species) but a limited number of Rhodophyta (1 of 5 species) and Chlorophyta (4 of 11 species), and all those RDRs formed clades that are distinct from land plants. We observed three distinct monophyletic groups (RDR1/2, RDR3 and RDR6) in plants, with RDR6 sister to RDR1 and RDR2, and RDR3 sister to RDR6 and RDR1/RDR2 (i.e., a tree topology of RDR3,(RDR6,(RDR2,RDR1))), in contrast with

recent publications reporting just two groups (RDR1/2 and RDR6) (You et al. 2017; S. Wang et al. 2021).

We observed that the Arabidopsis RDR3/4/5 proteins, encoded by tandemly duplicated genes, fell into the same (RDR3) clade, consistent with an earlier report (Hua et al. 2021). For simplicity, we refer to proteins in this clade collectively as RDR3 (Supplemental table 2). The *RDR3* genes likely emerged before the divergence of green and red algae, as their orthologs were detected in fungi and/or Metazoa. Notably, an *RDR3* copy is absent in all the eight gymnosperms, with an evolutionary pattern of (N-S,(M,E)) (N-S, non-seed land plants; E, eudicots; M, monocots) (Figure 1). There is only limited evidence for the function of *RDR3* – no compelling or complete data from Arabidopsis and only overexpression work (prone to artefacts and indirect effects) in rice (Jha et al. 2021). Thus, we hypothesize that the *RDR3* copies were lost in the gymnosperms, with as-yet unknown roles in earlier lineages and angiosperms.

Sister to RDR6, RDR1 and RDR2 likely diverged in a common ancestor of seed plants, with a tree topology of (N-S,((G1,(M1,E1)),(G2,(M2,E2)))) (N-S, non-seed land plant species; G, gymnosperms; E, eudicots; M, monocots)((Figure 1). Thus, we hypothesize that an ancestral *RDR* gene duplicated and gave rise to *RDR1* and *RDR2* that specialized in generating double-strand RNA molecules viral or heterochromatic ssRNA templates (Herr et al. 2005; Garcia-Ruiz et al. 2010). We found a cluster of proteins, from outgroup and green algae species, on a branch distinct to RDR6 and RDR1/2. We found two RDRs from the green algal species *Chara braunii* close to RDR6 and close to the RDR1/2 split, suggesting the presence of a proto-RDR6 and a proto-RDR1/2 and an early emergence of *RDR6 and RDR1/2* genes in genomes of land plants species. In land plants, all genomes encode RDR6 proteins with a species tree topology of (N-S,(G,(M, E))). RDR6 is the only major RDR group that is broadly present in the plant lineages. This is consistent with the broad presence of 21-nt *trans*-acting siRNAs (tasiRNAs) in land plants and the emergence of 21-nt and 24-nt reproductive phasiRNAs, which likely emerged in the angiosperms (Liu et al. 2020). These data in turn suggest that prior to the origin of angiosperms, RDR6 specialized in tasiRNA and

hc-siRNA pathways and became involved in the reproductive phasiRNA pathways in the angiosperms.

The origins of the DCL3/DCL5 divergence in angiosperm species

DCL proteins are composed of seven distinct protein domains (Supplemental Figure 1). Members of the DCL protein family are involved in all sRNA biogenesis pathways. Functional studies show that DCL1 catalyzes the biogenesis of miRNAs (Hiraguri et al. 2005) while DCL2 produces 22-nt siRNAs which are involved in antiviral defenses, transgene silencing, and other biological processes (Gasciolli et al. 2005; Ding and Voinnet 2007; Parent et al. 2015; Taochy et al. 2017; Wu et al. 2017; Wang et al. 2018; Jia et al. 2020). DCL3/4/5 are involved in the biogenesis of siRNAs (Borges and Martienssen 2015), but in different pathways; DCL4 and DCL5 process phasiRNAs from long Pol II products while DCL3 processes hc-siRNAs from short Pol IV products (Table 1). Recent work from our group (Patel et al. 2021) analyzed the divergence of DCL5 in several monocot species. However, the origin and evolutionary relationships of DCL3 and DCL5 beyond monocots have remained largely unclear.

To gain insights into the evolution of DCL3 and DCL5, we constructed a Maximum likelihood tree of 1,098 DCL proteins (including 25 non-plant DCLs) and rooted it at the branch that separates Metazoan, fungal and algal species from land plants. Isolated from others, we found Rhodophyta, Chlorophyta and other outgroup species in one phylogenetic group that is separated from land plants (Figure 2A; Supplemental data 2; Supplemental table 2). In sum, we identified four clades with a topology of (DCL1,(DCL3/5,(DCL4,DCL2))). We found two monophyletic groups, DCL1 and DCL4, present in all land plants (Figure 2A) with a species tree topology of (N-S,(G,(M, E))). Our results indicate an early divergence of these DCL proteins, implying a functional conservation of DCL1 and DCL4 in land plants, and suggesting the existence of both miRNA and 21-nt tasiRNA/phasiRNA pathways in ancient land plants. In support of this notion, previous phylogenetic studies had described the presence of DCL1 and DCL4 in bryophytes (You et al. 2017; S. Wang et al. 2021); others had detected the presence of miRNAs and 21-nt tasiRNAs in *Physcomitrella patens* (Arazi et al. 2005; Axtell et al. 2006; Talmor-Neiman et al. 2006; Axtell et al. 2007). The early emergence of DCL4 and

RDR6 (mentioned above) in land plants is consistent with previous reports demonstrating the presence of an active miR390-triggered *TAS3-ARF* regulatory module producing 21-nt phasiRNAs in *P. patens* and later lineages (Axtell et al. 2006; Axtell et al. 2007; Arif et al. 2012; Cho et al. 2012; Xia et al. 2017; Liu et al. 2020). The *DCL2* copies were only detected in seed plants following an evolutionary pattern of (G,(M, E)). This is consistent with recent studies showing the absence of a *DCL2* gene in genomes of the lycophyte *Selaginella moellendorffii*, and two moss species (*Physcomitrella patens* and *Sphagnum fallax*) (Axtell 2009; You et al. 2017). Together, these data suggest that the DCL2 and potentially the DCL2-derived 22-nt siRNAs emerged in seed plants.

DCL5, initially named DCL3b, catalyzes the biogenesis of 24-nt reproductive phasiRNAs (Song et al. 2012; Arikit et al. 2013; Teng et al. 2020; Patel et al. 2021). 24-nt reproductive phasiRNAs were found to peak in abundance at the meiotic stage of anthers (in rice and maize) and were initially described as grass-specific (Zhai et al. 2015; Fei et al. 2016). DCL5 was described as monocot-specific and only present in *Dioscorea* and more recently diverged monocot lineages, although DCL5 appeared to have been lost independently in some orders (Patel et al. 2021). We previously hypothesized that DCL5 emerged from DCL3 via WGD or a *DCL3* gene duplication some time before the diversification of *Dioscorea,* and became specialized for the biogenesis of 24-nt reproductive phasiRNAs by neofunctionalization (Patel et al. 2021). Although the genomes of eudicots do not encode the canonical DCL5 protein, it is now known that these genomes also generate 24-nt reproductive phasiRNAs, likely generated by DCL3 since no other plant DCL specializes in producing 24-nt small RNAs and there are no other candidate DCLs encoded in these genomes (Xia et al. 2019; Patel et al. 2021). Our data suggest that DCL5 did not emerge from DCL3 after the divergence between monocots and eudicots, but instead that eudicot genomes encode an ortholog to the monocot DCL5 rather than the monocot DCL3, and the monocot DCL3 clade diverged. Thus, the evolutionary pattern of DCL3 and DCL5 can be summarized as (N-S, (G,(M1,(M2,E2)))), where M1 refers to DCL3 and M2 and E2 refer to DCL5 (Figure 2A and 2B). We found that among monocots, DCL5 orthologs were present in *Zostera marina* and more recent monocot lineages although a few species

lost their *DCL5* copy or were absent from the assembly. In Figure 2B, the basal-most angiosperms, *Amborella trichopoda* (Amborellales) and *Nymphaea colorata* (Nymphaeales), do not cluster with monocots and eudicots, suggesting that DCL5 emerged some time before the diversification of the core angiosperms. Although there is only one *24-PHAS* locus, the production of 24-nt phasiRNAs has been described in *Amborella trichopoda* (Xia et al. 2019) in which we found only one *DCL3* copy (Figure 2B). In *A. trichopoda*, we hypothesize that DCL3 has dual functions to process distinct substrates for the biogenesis of either 24-nt phasiRNAs or 24-nt hc-siRNAs, a dual functionality apparently retained in DCL3 of many extant eudicots (Xia et al. 2019). In monocots, two paralogs have undergone subfunctionalization resulting in functional divergence of DCL3 and DCL5 to function in the 24-nt hc-siRNA and 24-nt phasiRNA pathways, respectively. The maize loss-of-function *dcl5* mutant fails to produce 24-nt reproductive phasiRNAs but not hc-siRNAs (Teng et al. 2020), demonstrating its specialized function. The conflict that we observe relative to our prior work (Patel et al. 2021) that hypothesized the emergence of DCL5 only in the monocots is that the eudicot DCL3 is closer to the monocot DCL5 than the monocot DCL3 copy. That work and other earlier studies were performed on a limited number of closely related species (Xia et al. 2019; Bélanger et al. 2020; Patel et al. 2021; Pokhrel et al. 2021).

Perhaps the most parsimonious hypothesis to explain these observations is that the earliest monocot ancestor apparently gained a second copy of DCL3, and evolved this into a specialized hc-siRNA-generating DCL (still called DCL3). Yet, despite apparent functional equivalence, particularly known for Arabidopsis (that lacks the 24-nt phasiRNAs and thus also only makes hc-siRNAs), the monocot DCL3 has diverged in sequence relative to the monocot DCL5. With this specialized *DCL3* present in the monocot genome, the eudicot-like *DCL* copy (now called *DCL5*) subfunctionalized and specialized to generate 24-nt phasiRNAs. In parallel, the only copy of DCL3 retained in the eudicots (and perhaps magnoliids, as evidenced by *Cinnamomum kanehirae*, “Cin.ka” in Figure 2) likely retained the dual functions of producing 24-nt phasiRNAs and 24-nt hc-siRNAs. While alternative hypotheses could be constructed to explain DCL3 evolution (such as gene duplication prior to the monocot-eudicot split, with a copy lost in

eudicots), the hypothesis explained here seems both the most simple and consistent with the data.

An outcome of this analysis is that the DCL3 and DCL5 protein names may not accurately reflect their evolutionary history. However, monocot DCL3 and DCL5 are clearly functionally divergent as they act in distinct pathways processing hc-siRNA versus phasiRNA, involved in different developmental processes, using distinct substrate RNAs, and presumably different subcellular localization (nucleus versus cytoplasm). For historical reasons, we recommend maintaining the DCL3/DCL5 nomenclature emphasizing their functional differences, although the DCL3 name for eudicots with both 24-nt hc-siRNAs and phasiRNAs may ignore a dual functionality. Future biochemical studies that test this functionality across diverse eudicots may find it helpful to revisit the nomenclature.

Sequence divergence in specific domains underlies the functional divergence of DCL3, DCL4 and DCL5 proteins

To understand the molecular basis of the functional differences among the DCL3, DCL4 and DCL5 proteins, we examined the conservation and divergence in their amino acid sequences. These proteins are specialized to process dsRNA substrates in single or regular duplex increments producing sRNA duplexes of 21-nt or 24-nt. The evolutionary processes driving these DCL functions are not well examined. Assuming that shared and divergent catalytic activities result from the evolution of specific amino acids, we created several groups of sequences: amino acids in DCL3, DCL4 and DCL5 conserved at 95% among all monocot protein sequences; amino acids that are exclusive to DCL3/DCL5 or DCL4/DCL5; and strings of amino acids unique to specific proteins. Then, we visualized the position of these amino acids on the rice DCL5 protein structure predicted by AlphaFold (Jumper et al. 2021). In rice, DCL3, DCL4 and DCL5 proteins are of a similar length (1651 amino acids for DCL3; 1657 amino acids for DCL4; 1637 amino acids for DCL5) and all contain seven protein domains. Figure 3A shows the predicted rice DCL5 structure with known protein domains annotated. To assess the validity of the predicted rice DCL5 AlphaFold protein structure, the overlap of the alpha helices and beta sheets were compared by eye with the same domains

recently structurally elucidated by cryo-EM for the Arabidopsis DCL3 protein (Supplemental Figure 3) (Q. Wang et al. 2021).

From the grouping of conserved versus divergent amino acids, we identified a total of 167 amino acid positions conserved among monocot proteins for DCL3, DCL4 and DCL5 . Among these, we found 86 amino acids that are conserved among DCL3, DCL4 and DCL5; 35 amino acids conserved between DCL4 and DCL5; and 46 amino acids between DCL3 and DCL5. We found that most conserved amino acids are located in the inner core of the DCL5 structure (Figure 3B) known to bind the dsRNA molecule (Q. Wang et al. 2021). The amino acids conserved between the three DCL proteins (in green) are primarily located in the helicase, RNAse IIIa, and RNAse IIIb domains (Figure 3B), suggesting these regions are the key functional domains of the DCL proteins and have been maintained by natural selection. We also observed that amino acids conserved only between DCL3/DCL5 or DCL4/DCL5 were located in distinct protein domains. Most amino acids conserved between DCL3/DCL5 (Figure 3B; in yellow) were in the PAZ, RNAse IIIa, and RNAse IIIb domains, while most of those conserved between DCL5/DCL4 (Figure 3b; in black) were in the PAZ and dsRBDa domains. Absent in DCL4 are amino acid regions conserved between DCL3/DCL5 proteins (Figure 3C) that are located directly at two junctions between the PAZ domain and adjacent domains (which are 7 and 8 amino acids, in red) or on folded regions near the PAZ domain (which are 6, 5 and 10 amino acids, in deep olive). The distance between the PAZ and RNase III catalytic centers has been described to determine the sRNA product size (MacRae et al. 2006; MacRae et al. 2007); thus, the presence (DCL3/DCL5) or absence (DCL4) of these regions may underlie the specific roles of DCL3/DCL5 and DCL4 in processing 24-nt and 21-nt siRNAs, respectively.

Comprehensive annotation of AGO *genes in plants*

AGO proteins form RNA-induced silencing complexes (RISCs) with sRNAs to function in many gene silencing pathways. The number of AGO proteins has greatly expanded in gymnosperms (*Picea abies*, 9 AGOs (Z. Li et al. 2021)), eudicots (*A. thaliana*; 10 AGOs; (Vaucheret 2008; Zhang et al. 2015)) and monocots (*O. sativa*, 19 AGOs (Kapoor et al. 2008; Zhang et al. 2015)) compared to mosses (*P. patens*, 6 AGOs

(Axtell 2009; Zhang et al. 2015)) and ferns (*Azolla filiculoides*, 6 AGOs (Z. Li et al. 2021)). In angiosperms, AGO proteins form three major phylogenetic clades (Vaucheret 2008; Zhang et al. 2015) known to have distinct functions in binding to different groups of sRNAs. However, it is still unclear how AGO proteins evolved in seed plants and, more specifically, in monocots, in which the numbers of *AGO* genes are generally larger. Thus, we constructed a Maximum likelihood tree of 2,991 AGO proteins (including 85 non-plant AGOs) (Supplemental Figure 4; Supplemental data 3; Supplemental table 2). We found three major clades of AGO proteins, i.e., the AGO4/6/8/9, AGO2/7 and AGO1/5/10/17/18 clades (Figure 4 and 5; Supplemental data 3). We found a cluster of non-Viridiplantae organisms (Metazoan, Fungi and Rhodophyta) with Chlorophyta species (green algae; Viridiplantae) that clearly evolved separately from land plants (Supplemental Figure 4), consistent with an early divergence and specialization of AGOs in land plants. Consistent with the prior reports, the number of *AGO* copies per genome greatly varied between plant lineages, ranging on average from 1.2 copies in Chlorophyta, 9.3 in Bryophyta, 5.0 in Lycopodiopsida, 7.0 in Polypodiopsida, 9.3 in Acrogymnospermae, to 20.5 in Liliopsida and 14.5 in Pentapetalae for Mesangiospermae species (Table 2). A recent publication described a phylogenetic analysis of 2,958 AGO proteins from 244 plant species (Z. Li et al. 2021). Although we analyzed 50 fewer plant species, we identified a larger number of AGO proteins than (Z. Li et al. 2021), suggesting that our work found a more complete set of AGO proteins and may be more comprehensive. For example, we found many more AGOs in *Hordeum vulgare*, *Sorghum bicolor*, *Zea mays* and *Triticum aestivum*, with 18, 18, 21, and 71 AGOs, respectively, compared to 12, 15, 17, and 40 AGOs identified by Li et al. (2021). Such differences were likely due to (i) the source of genome annotation, in our case prioritizing EnsemblGenomes and Phytozome databases, rather than NCBI, (ii) the sensitivity of the method we used to identify orthologous AGOs (e.g., the use of OrthoFinder), and (iii) the addition of recently sequenced large monocot genomes (e.g., 22 AGOs in *Secale cereale*).

Distinct patterns of AGO4 expansion in monocots versus eudicots

Members of the AGO4/6/8/9 clade have been described as effectors of 24-nt hc-siRNAs and are known to function in transcriptional gene regulation (Zhang et al. 2015). Our phylogenetic tree showed that there were small numbers of AGO-like proteins in the moss, fern and gymnosperm lineages, whereas in the flowering plants, the AGO family expanded and diverged extensively, resulting in groups of AGO4 and AGO6 (Figure 4A; Supplemental data 3), with a evolutionary pattern of (N-S,(G,((M,E),(M,E)))) for (AGO-like,(AGO-like,((AGO6,AGO6),(AGO4,AGO4)))). As a side note, AGO protein names were based largely on initial and historical phylogenetic studies of Arabidopsis, based on which distinct names (AGO4/6/8/9) were assigned to four related AGO proteins. These names were then applied to rice, maize, and other species (Zhang et al. 2015), in some cases confusing relationships. Based on our analyses, we suggest adopting a gene nomenclature using the plant lineage group to designate the origin of AGO paralogs, for instance, Brassicales-AGO4-i, Brassicales-AGO4-ii, and Brassicales-AGO4-iii for AGO4 paralogs. Our analyses of the AGO6 and AGO4 proteins suggest that the AGO clade previously known as AGO4/6/8/9 should be renamed as AGO4/6, and the AGO8 and AGO9 names used only for Arabidopsis, in which it is too late to correct the naming.

Our results indicate that AGO6 and AGO4 diverged in a common ancestor of core angiosperms. Indeed, basal angiosperms (Figure 4A); *Amborella trichopoda* (Amborellales) and *Nymphaea colorata* (Nymphaeales), two basal-most angiosperms, group together before the split of AGO6 and AGO4 while *Cinnamomum kanehirae* (Magnoliid; Laurales) and *Aquilegia coerulea* (basal eudicot; Ranunculales) have members of AGO proteins in both clades, like the eudicots and core angiosperms (Figure 4A). The AGO6 proteins formed a monophyletic clade including all core angiosperms, such that the protein tree mirrored that of the species tree (Figure 4A), suggesting that AGO6 function is highly conserved within angiosperms. This is consistent with functional studies performed in Arabidopsis demonstrating that AGO4 (AGO4 and AGO9) and AGO6 are partially redundant, as their functions are conserved in the RdDM pathway, but the encoding genes have distinct spatial and temporal

activities (Havecker et al. 2010; Duan et al. 2015). Both AGO4/6 bond to 5' adenosine 24-nucleotide sRNAs, but have preferences for sRNA associated with different heterochromatin loci determined by the coincident expression of the AGO and sRNA-generating loci rather than functional difference between AGO4 and AGO6 (Havecker et al. 2010). Similarly, physical co-localization of AGO4 and AGO6 with Pol II and Pol V is determinant in the function of these AGOs in Arabidopsis (Duan et al. 2015). Having said that, the AGO4 proteins display more complex evolutionary patterns (Figure 4A) such that the protein tree does not reflect that of the species tree, suggesting that the *AGO4* genes have undergone extensive duplication and possibly neofunctionalization/subfunctionalization.

Within monocots, AGO4 exhibited distinct evolutionary patterns in Poaceae and non-Poaceae species. We observed intra-species *AGO4* duplication events in several species within the Zosteraceae (*Zostera marina*), Asparagaceae (*Asparagus officinalis*), Orchidaceae (*Dendrobium catenatum*, *Phalaenopsis equestris*), Bromeliaceae (*Ananas comosus*) and Cyperaceae (*Carex littledalei*) families, resulting in two *AGO4* copies per genome. In the Poales, we observed that a Bromeliaceae (*Ananas comosus*) and Cyperaceae (*Carex littledalei*) species each have two tandemly duplicated *AGO4* copies clustered species-by-species on independent branches to Poaceae suggesting independent gene duplication events (Figure 4A). In contrast, three *AGO4* copies were detected in almost every Poaceae species, forming three monophyletic groups (Figure 4A), which may be the result of a whole genome duplication followed by a tandem duplication of one *AGO4* gene copy in the common ancestor of the Poaceae family. This includes one more highly diverged clade (Poaceae AGO4-i, Figure 4A); future functional analyses and assessments of redundant and unique functions may be informative. We hypothesize that the different number of gene duplication events observed in the AGO4 and AGO6 clades is likely to be associated with the divergence of sRNA pathways in which they are involved. For example, it has been suggested that AGO4 and AGO6 are specifically involved in the RdDM pathway, whereas AGO9 and AGO104 (in the group Poaceae-AGO4-iii) proteins are believed to play important roles in regulating plant reproductive development in Arabidopsis and maize (Olmedo-Monfil et al. 2010; Singh et al. 2011), respectively. Together, our current phylogenetic data,

and the published functional data, suggest multiple gene duplication and neofunctionalization events specific to the *AGO4* genes in monocots.

The evolution of AGO4 in eudicots is even more complex since the AGO4 group expanded several times independently in the genomes of various species or taxonomic groups. For example, divided into two branches, AGOs from Fagales species form a clade containing genes from *Carpinus*, *Fagus*, *Juglans* and *Quercus* species, beside other rosid species. A second example is the evolution of AGO4 in the Brassicales. We observed three AGO4 subgroups (named Brassicales AGO4-i, AGO4-ii and AGO4-iii) that form a single larger clade distinct from other species in the Malvids group (Figure 4A). Based on the annotation of Arabidopsis genes, the *AGO4* genes in Brassicales AGO4-ii (represented by the Arabidopsis *AGO8*, AT5G21030) appears to be duplicated from the paralog in Brassicales-AGO4-i (represented by the Arabidopsis *AGO4*, *AT2G27040*) following a WGD, while a tandem duplication may have been the mechanism to generate the AGO4 paralog in the group AGO4-iii (represented by the Arabidopsis *AGO9, AT5G21150*); all species are represented in Brassicales AGO4-i and AGO4-iii, while several species lost their copy in the AGO4-ii subgroup from which the *AGO8* gene (AT5G21030) has been suggested as pseudogene in Arabidopsis (Mallory and Vaucheret 2010).

AGO2/7 emerged in the common ancestor of ferns and seed plants and diverged in seed plants

Members of the AGO2/7 clade are effectors of miRNAs and known to function in antiviral defenses and post-transcriptional gene regulation (Harvey et al. 2011; Zhang et al. 2015). Our phylogenetic tree shows a group of AGO2/7-like proteins present in fern that then diverged and expanded in seed plants, resulting in two groups of AGO2 and AGO7 (Figure 4B; Supplemental table 2). We did not identify any AGO2/7-like protein encoded in the genomes of four moss species suggesting that the AGO2/7 clade emerged in the common ancestor of vascular plants and then diversified in seed plants. We infer that AGO7 emerged in seed plants, as we identified one AGO7 ortholog in most gymnosperms, and then observed a monophyletic group of AGO7 covering all flowering plants (Xia et al. 2017). AGO7 catalyzes the miR390-directed cleavage of

TAS3 transcripts, yielding 21-nt tasiRNAs (aka tasiARFs) regulating ARF family transcription factors in angiosperms (Montgomery et al. 2008). Previous studies had reported a miR390-TAS3-ARF regulatory module acting in the bryophyte *Physcomitrella patens* and some later lineages (Axtell et al. 2006; Axtell et al. 2007; Arif et al. 2012; Cho et al. 2012; Liu et al. 2020). Thus, we hypothesize that an AGO protein other than AGO7 is involved in the biogenesis of tasiARFs in bryophytes, and the miR390-triggered AGO7-catalyzed TAS3-ARF regulatory module emerged in a common ancestor of gymnosperm and angiosperm species.

AGO2 likely emerged in the common ancestor of seed plants, as genomes of the gymnosperm, basal angiosperm and core angiosperm lineages encode AGO2, but the lycophytes and bryophyte genomes do not. In monocots, subgroups of plant species appear to have lost their *AGO2* gene copy as genomes of Alismatales (Araceae: *Spirodela polyrhiza*; Zosteraceae: *Zostera marina*), Asparagales (Orchidaceae: *Apostasia shenzhenica*, *Dendrobium catenatum*, *Phalaenopsis equestris*), Dioscoreales (Dioscoreaceae: *Dioscorea alata*, *Dioscorea rotundata*), Zingiberales (Musaceae: *Musa acuminata*, *Musa balbisiana*) and Poales (Cyperaceae: *Carex littledalei*) species do not encode AGO2 protein (Figure 4B). The *AGO2* gene was likely duplicated in a common ancestor of the Poaceae species, resulting in two *AGO2* clades (Figure 4B). One group of *AGO2* genes (Poaceae-AGO2-i) was preserved in most Poaceae species. Among these, some *AGO2* genes were specifically duplicated in the *Oryza* species, forming an Oryzoideae-specific clade. A second group of AGO2 proteins (Poaceae-AGO2-ii) was lost in genomes of PACMAD species and Oryzoideae (BOP clade) species but maintained in Pooideae (BOP clade) genomes.

Several eudicot genomes have multiple AGO2 paralogs; AGO2 proteins do not cluster in monophyletic groups suggesting that they did not diversify from a gene duplication event in a common ancestor of eudicots. As an example, the phylogenetic tree in Figure 4B shows a complex and ambiguous evolution of *AGO2* genes in the genomes of Asterids. We observed three clusters of AGO2 proteins grouping all species from the *Solanum* genus but distinct to other species in the Solanaceae family and Solanales order, consistent with the recent publication about Solanaceae AGO2 protein (Baldrich

et al. 2021). Similarly, we observed two AGO2 proteins in each of four Asterales (*Helianthus annuus*, *Lactuca saligna*, *Lactuca sativa* and *Mikania micrantha*) and one Apiales (*Daucus carota*). All these AGO2 proteins are on the same branch and grouped by genus. AGO2 from the two *Lactuca* species are grouped together per species. Also, all these AGO2 resulted from tandem gene duplication events. Taken together, our observations suggest that duplication events arose genus-by-genus in eudicot genomes but not in a common ancestor. Together, the evolution of *AGO2* genes in the Poaceae and eudicots exhibit complex and species- and/or genus-specific patterns, which might be consequences of co-evolution with various bacterial/viral pathogens.

Expansion of the AGO1/5/10/18 clade in monocots

AGO1 and AGO10 proteins are the canonical AGOs known to interact with miRNAs to regulate plant development (Zhang et al. 2015). Our phylogenetic tree shows a proto-AGO1/5/10-like cluster that groups ferns and gymnosperms (Figure 5A). Curiously, all AGO proteins in mosses are grouped with other non-flowering species on a branch just adjacent to the AGO10 cluster, rather than with the proto-AGO1/5/10-like group (Figure 5A). Assuming that the proto-AGO1/5/10-like group contains the "original ancestral" AGO protein of the AGO1/5/10/18 clade, of which the coding gene was then duplicated and diversified, our phylogenetic tree suggests that mosses lost the *proto-AGO1/5/10*-like copy but retained an *AGO1/10*-like copy, and genomes of non-flowering plants lost the *AGO1* and *AGO5* (and *AGO18*) copies.

Our phylogenetic tree shows that AGO1 and AGO10 diverged from the same branch. As mentioned above, an AGO10-like clade grouping moss, fern and gymnosperm species are on a branch just adjacent to the AGO10 cluster rather than forming an AGO1/10-like clade before the divergence of AGO1 and AGO10 (Figure 5A). Furthermore, genomes of gymnosperm and angiosperm species encode an AGO10 protein while AGO1 is only present in angiosperm species. A recent phylogenetic analysis of plant AGO proteins detected the same features (Z. Li et al. 2021). These observations suggest that the moss/fern group had a single AGO1/10-like protein that loads all microRNAs, while gymnosperms had two – one inherited from the moss/fern group and a more diverged copy. Angiosperms may have inherited the more diverged

copy, and a duplication event within the angiosperms gave rise to AGO1, which became the key AGO for microRNA functions. As the evolution of AGO10 was recently well described, including the split into AGO10a and AGO10b clades in the angiosperms (Z. Li et al. 2021), we focus on the evolution of AGO1 in monocots. Our phylogenetic tree indicates that a duplication event occurred early after the monocot/eudicot split, resulting in two groups (Figure 6A). Then, another duplication event of each group occurred in a common ancestor of the Poaceae to generate two sub-groups, resulting in four groups, called Poaceae-AGO1-i to AGO1-iv (Figure 6A). Further gene duplication or loss events arose in specific lineages at one or another branch. First, the group AGO1-i represents all Poaceae species but maize and sorghum. Second, all PACMAD species are absent from AGO1-iii. Finally, *AGO1-iv* was duplicated in Pooideae and PACMAD species. That is, duplicated *AGO1-iv* copies are observed in all PACMAD species. In Pooideae, duplicated *AGO1-iv* copies are present in genomes of Triticeae species but absent from Brachypodieae genomes. As the canonical effector(s) of miRNAs, the expansion of AGO1 might be associated with the evolution of specialized sRNA pathways involving miRNAs (e.g., perhaps the reproductive phasiRNA pathways, which seem to have diversified particularly in grass species, relative to other monocots).

AGO18 has been described as a Poaceae-specific protein (Zhang et al. 2015) and proposed as a candidate effector of 24-nt reproductive phasiRNAs in anther development based on gene co-expression analyses (Zhai et al. 2015; Fei et al. 2016; Bélanger et al. 2020). As expected, we detected at least one copy of *AGO18* in all Poaceae genomes. Interestingly, our phylogenetic analysis detected an AGO18 protein in *Carex littledalei* (Poales; Cyperaceae) but not in *Ananas comosus* (Poales; Bromeliaceae), indicating that AGO18 emerged in a common ancestor of Cyperaceae and Poaceae (Figure 5B). The AGO18 clade is curious because of its basal location on phylogenetic trees but emergence only in a few families of monocots; with the current data, it seems challenging to determine if AGO18 emerged in Cyperaceae species or was lost in other lineages. AGO17, found in the *Oryza ssp.*, *Aegilops ssp.*, *Eragrostis tef*, *Miscanthus sinensis*, *Thinopyrum intermedium* and *Triticum ssp.*, is similarly unusual. While some studies have identified functions (Wu et al. 2015; Sun et al. 2018;

Sun et al. 2019), these lack commonalities suggesting that a better understanding and interpretation of the role of AGO18 will require more experimentation.

The AGO5 proteins were shown to be angiosperm specific (Z. Li et al. 2021 a). The number of AGO5 proteins greatly expanded in genomes of Poaceae species, and the evolutionary pattern of this AGO is more complicated than that of AGO1, AGO2 or AGO4. Poales AGO5 proteins diverged from *Musa* and other monocot lineages before expanding in Poaceae species, resulting in three major groups, which we have named Poaceae-AGO5-i to AGO5-iii (Figure 6B). Some lineages have extra *AGO5* copies in the Poaceae-AGO5-i clade. We observe that AGO5-i split two times in a common ancestor of Poaceae resulting in three branches where (i) one branch is represented by Oryzoideae species only, (ii) the second branch is represented by Pooideae species only and (iii) the third branch represent all Poaceae species (Figure 6B). That is, AGO5-i demonstrates signatures of duplication consistent with a gain of a copy in Oryzoideae and Pooideae lineages. Similarly, we observed two *AGO5* copies in Oryzoideae species of Poaceae-AGO5-ii while only one in all other Poaceae, suggesting another single gain of an *AGO5* copy in genomes of the Oryzoideae group (Figure 6B). AGO5 in the Poaceae-AGO5-iii also expanded but followed a distinct evolutionary pattern as based on their arrangement in *Oryza sativa* ssp *japonica* genome they appear to be tandem duplicates. Interestingly, one of the duplicated AGO5-iii genes encodes the rice AGO5c protein, also called MEIOSIS ARRESTED AT LEPTOTENE 1 (MEL1), and is known as a binding partner of 21-nt reproductive phasiRNAs and essential for the progression of meiosis in anthers (Nonomura et al. 2007; Komiya et al. 2014). Similarly, in maize, *AGO5c* has recently been identified as the causal gene of the classic mutant *male sterile 28* (*ms28*) and is essential for anther (tapetum) development and male fertility (Y. Li et al. 2021). In separate work, both AGO5b and AGO5c–named as MALE-ASSOCIATED ARGONAUTE-1 (MAGO1) and -2 (MAGO2), respectively, in that work–load 21-nt phasiRNAs in pre-meiotic anthers to regulate retrotransposons (Lee et al. 2021). These published data suggest a conserved role of AGO5 proteins in the premeiotic, 21-nt phasiRNA pathway. Interestingly, MAGO1 is a member of the AGO5-iii clade while MAGO2 is in the AGO5-ii clade, suggesting some level of redundancy of AGO5 proteins associated with 21-nt phasiRNAs in male reproductive development.

In summary, our results suggest that (i) green algae genomes do not encode any AGO protein in this clade, known as the effector of miRNAs, (ii) AGO10 emerged in seed plants and then expanded in two clades in flowering plants, (iii) only angiosperm genomes encode AGO5, and (iv) AGO18 is present only in monocots. Moreover, our data indicate the evolutionary complexity of AGO1, AGO5 and AGO18 in monocots by demonstrating that *AGO* gene copy number has increased in a common ancestor of Poaceae and then followed by lineage-specific gene gain or loss, retention, or tandem duplication events. Since AGO1, AGO5, and AGO18 are the only three clades that have been shown to be involved in reproductive phasiRNA pathways, their expansion in the grass lineage may have been driven by a need to form RISCs with miR2118 or miR2275, to trigger production of either 21- or 24-nt reproductive phasiRNA production, or to load the phasiRNAs and act at their targets.

Phylogenetic analysis of accessory proteins involved in sRNA biogenesis

To further understand the evolution of protein complexes involved in sRNA biogenesis and modification, we performed phylogenetic analyses of the DRB and SE (putative protein partners of DCLs) and SGS3 (putative partners of RDRs) proteins. Although not directly involved in sRNA biogenesis, the HUA ENHANCER 1 (HEN1) protein functions to add a methyl group that protects and stabilizes sRNAs (Z. Yang et al. 2006). Maximum likelihood trees of 1,903 DRB (Supplemental Figure 5), 286 HEN1 (Supplemental Figure 6), 486 SE (Supplemental Figure 7) and 465 SGS3 (Supplemental Figure 8) proteins were constructed and rooted (Supplemental data 1; Supplemental table 2).

DCL proteins function with DRB and SE cofactors. Previous research showed that DRB1 (also named HYL1) and SE interact with DCL1, while DRB4 interacts with DCL4 in Arabidopsis (Hiraguri et al. 2005; Nakazawa et al. 2007; Curtin et al. 2008; Fukudome and Fukuhara 2017). Our phylogenetic analysis identifies five distinct DRB protein clades, as well as many of the terminal branches. Compared to the other families that we analyzed, some branches are quite long, in particular, the length of the branch separating DRB3 from other DRB proteins as well as the length separating monocots and eudicots in DRB3. Although some DRB proteins had been detected in outgroup

species, no DRB proteins had been detected in Chlorophyta species suggesting that DRB independently emerged in the common ancestor of land plants or was lost in the Chlorophyta lineage. Rooted with non-plant outgroup species, the clade named DRB3 clusters on a separate branch to other DRB proteins and includes only flowering plants. The evolution of DRB1, a canonical cofactor of DCL1, is monophyletic as it is present in moss and descendent lineages (Supplemental Figure 5). This observation is consistent with the absence of DCL1 protein in Rhodophyta and Chlorophyta species, suggesting that DCL1 and DRB1 emerged simultaneously to enable miRNA biogenesis. Producing 21-nt phasiRNAs, the miR390-TAS3-ARF regulatory module requires DCL4 and, presumably, DRB4 protein as a cofactor; however, DRB4 has been found only in flowering plants, while tasiRNAs emerged in mosses (Arazi et al. 2005; Axtell et al. 2006; Talmor-Neiman et al. 2006; Axtell et al. 2007). Besides DRB1, only the clade DRB2-iii was observed in all plant lineages. DRB4 and DRB2 descended from the same ancestral gene (Supplemental Figure 5), and thus DRB4 and DRB2 may have diverged from a Dicer-associated DRB protein. Assuming that DCL4 needs a DRB protein partner to process dsRNAs into phasiRNAs, we hypothesize that DRB2-iii could be a cofactor of DCL4 in early land plant species.

Present in most outgroup species, including Rhodophyta, our phylogenetic analysis shows that SE protein is encoded in the genome of Chlorophyta and land plants (Supplemental Figure 7). The Arabidopsis genome encodes only one SE protein, described as a partner to DCL1 in the biogenesis of miRNA (L. Yang et al. 2006). Recent work has demonstrated a dual role of SE, in both miRNA processing and in RNA decay as part of the nuclear exosomal complex (Bajczyk et al. 2020); thus, it may be well conserved because it lacks exclusivity of particular RNAs in these activities. Interestingly, we observed that the SE protein family diversified in monocots sometime before the split of Poaceae and other Poales species. We identified three Poaceae groups, which we named SE-i, SE-ii and SE-iii (Supplemental Figure 7). Since these SE proteins diverged in a common ancestor of Poaceae, these three groups may have evolved separately and undergone subfunctionalization or neofunctionalization.

The HEN1 protein, encoding an RNA methyltransferase, is involved in sRNA modifications, and is conserved also in animals (Kirino and Mourelatos 2007). We observed that the HEN1 protein was not present in all outgroup species, including Rhodophyta and Chlorophyta, but one copy or more was found in all land plant species (Supplemental Figure 6). HEN1 has remained a single-copy gene in most species including the monocots, but shows modest expansion in some eudicot lineages. For example, we detected two *HEN1* copies in all Brassicaceae species but not in *Cleome violacea*, suggesting that HEN1 expanded in a common ancestor of Brassicaceae through a WGD (Supplemental Figure 6).

Finally, we examined the evolution of the SGS3 family, among which the Arabidopsis SGS3 protein is known to be required for tasiRNA biogenesis (Peragine et al. 2004). Consistent with the role of SGS3 in the plant-specific tasiRNA pathway, no SGS3 proteins were found in Metazoa species that we used as outgroups, nor in the Rhodophyta and Chlorophyta species, suggesting that SGS3 emerged in the genome of a common ancestor of land plants (Supplemental Figure 8). This observation is also consistent with the absence of RDR protein in the Metazoa outgroup, and most Rhodophyta and Chlorophyta species, suggesting that RDR and SGS3 emerged together, for their common role in tasiRNA biogenesis. Although most land plants have two *SGS3* copies per diploid genome, all proteins evolved from the same branch such that the protein tree mimicked that of the species tree, suggesting that no subfunctionalization or neofunctionalization evolutionary processes occurred in SGS3 paralogs before speciation. That is, there is no phylogenetic evidence indicating that SGS3 proteins could specialize in an sRNA biogenesis pathway by becoming a specific partner to various RDR proteins, like RDR2 or RDR6 in the hc-siRNA or phasiRNA pathways, specifically.

CONCLUDING REMARKS

Compared to the rapidly growing number of plant genomes that are sequenced, comprehensive evolutionary studies of proteins associated with sRNA biogenesis and function in plants remain limited. Two recent publications that studied the emergence of some of the proteins analyzed in this work across ancestral lineages of Embryophyta investigated a total of 36 (You et al. 2017) or 34 (S. Wang et al. 2021) species. Those studies uncovered the origin of AGO, DCL, HEN1, RDR, SE and other sRNA biogenesis proteins; however, both studies investigated a limited number of gymnosperm (1) and angiosperm (3) species. Other phylogenetic studies investigated the evolution of some of the proteins studied in this work (mainly AGO, DCL and RDR proteins) in angiosperms. However, these studies were performed on a small number of closely related species (Xia et al. 2019; Bélanger et al. 2020; Pokhrel et al. 2020; Patel et al. 2021) and thus a complete evolutionary history could not be elucidated. The evolution of the DCL proteins (DCL3/DCL5 in particular) and AGOs (AGO4/6 and AGO1/5/10, in particular) was not thoroughly analyzed. We detected multiple duplication events of *AGOs* that have been lost, retained or further duplicated in sub-groups, indicating complex patterns of AGO diversification in monocots. This also suggests that the conclusion of functional studies in model species such as Arabidopsis may not be easily translated to other groups of species. The number of plant genome assemblies keeps increasing rapidly, while bioinformatic programs and cheap computing resources facilitate evolutionary studies over tens to hundreds of species, from simple unicellular organisms like algaes to complex allopolyploid and large genome like wheat. Therefore, future studies may continue to refine our understanding of the evolution of proteins functioning in plant small RNA biology, while functional analysis and experimentation may discern the roles and contributions of different protein clades.

MATERIALS AND METHODS

Datasets Used in this Work

Proteome sequence files were downloaded from Ensembl Genomes, Phytozome, NCBI, Plaza, TreeGenes, FernBase, OrchidBase, Sweetpotato Genomics Resource and Sol Genomics Network. A detailed list of species analyzed in this work are provided in Supplemental Table 1, including the source and version of references used and the taxonomic classification. Species names were shortened to five-letter codes (i.e., “species codes”), which were used to annotate phylogenetic trees.

Construction and Visualization of the Species-Tree

The NCBI Taxonomy Browser (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) was used to build a species-tree and visualize the evolutionary relationship among the studied species. iTOL v4 (Letunic and Bork 2019) to visualize the species-tree.

Identification of Orthologs and Build Phylogenetic Trees

OrthoFinder v2.3.11 (Emms and Kelly 2015) was used to infer gene orthology among 219 species using the following parameters: -I 2 -S diamond -M msa -A mafft -T iqtree -os. To identify orthologous groups containing DCL, DRB, SE, HEN1, RDR, SGS3 and AGO protein families, reference protein sequences annotated in *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Hordeum vulgare*, *Oryza sativa*, *Physcomitrella patens*, *Picea abies*, *Solanum lycopersicum*, *Triticum aestivum*, *Volvox carteri* and *Zea mays* (reported by (Zhang et al. 2015; You et al. 2017; Bélanger et al. 2020) or available on UniProt database) were used. Protein sequences manually curated for presence of functional domains (detailed in Supplementary Figure 1) using either ScanProsite (Gattiker et al. 2002; Castro et al. 2006) or CDvist (Adebali et al. 2015). Multiple sequence alignments (MSAs) were performed using MAFFT v7.45 (Katoh and Standley 2013) with the following parameters: --maxiterate 1000 --localpair. Protein alignments were trimmed with trimAL v1.4.rev15 (Capella-Gutiérrez et al. 2009) using parameters optimized for each alignment. Trimmed alignments were used to infer maximum-likelihood phylogenetic trees using IQ-TREE v1.6.12 (Minh et al. 2013; Nguyen et al.

2015; Kalyaanamoorthy et al. 2017) with the following parameters: -alrt 1000 and -bb 1000. iTOL v4 (Letunic and Bork 2019) to draw and annotate consensus trees.

Phylogenetic trees in Newick format are provided in Supplementary Data 1-7. All proteins in the trees were annotated with species names, protein IDs and protein names (if known); e.g., Hor.vu-DCL5|HORVU.MOREX.r2.1HG0034750.1 represents the DCL5 protein in barley (*Hordeum vulgare*). A complete list of protein IDs and names for all species are provided in Supplementary Table 2.

All annotated proteins and trees can be consulted, allowing readers to retrieve the annotated protein ID from the 198 plant species and to visualize the phylogenetic analyses.

Comparative visualization of protein structures

Multiple sequence alignments were performed on protein sequences of interest using MAFFT v7.45 (Katoh and Standley 2013) with the following parameters: --maxiterate 1000 --localpair. Sequence alignments were first visualized using the SnapGene software (from Insightful Science; available at snapgene.com) to identify conserved amino acid residues(> 95% identity) between sequences. Then PyMol v2.4.2 (https://pymol.org/2/) was used to visualize the location and distribution of residues over the rice 3D protein structure predicted by AlphaFold (Jumper et al. 2021). Proteins were manually annotated to indicate key domains and the positions.

ACKNOWLEDGMENTS

We thank Danforth Center colleagues Elizabeth Kellogg and Yunqing Yu for helpful discussions, and Joanna Friesner for assistance with editing. This work was partly supported by a USDA National Institute of Food and Agriculture “BTT EAGER” award no. 2018–09058 to B.C.M.). Sébastien Bélanger was supported by a postdoctoral fellowship from the Natural Sciences and Engineering Council of Canada (NSERC).

CONTRIBUTIONS

SB, JZ, and BCM designed the analyses. SB performed the analyses. SB, JZ, and BCM interpreted the results. SB, JZ, and BCM wrote the manuscript.

BIBLIOGRAPHY

Adebali O, Ortega DR, Zhulin IB. 2015. CDvist: a webserver for identification and visualization of conserved domains in protein sequences. *Bioinformatics* 31:1475–1477.

Arazi T, Talmor-Neiman M, Stav R, Riese M, Huijser P, Baulcombe DC. 2005. Cloning and characterization of micro-RNAs from moss. *Plant J* 43:837–848.

Arif MA, Fattash I, Ma Z, Cho SH, Beike AK, Reski R, Axtell MJ, Frank W. 2012. DICER-LIKE3 Activity in Physcomitrella patens DICER-LIKE4 Mutants Causes Severe Developmental Dysfunction and Sterility. *Mol Plant* 5:1281–1294.

Arikit S, Zhai J, Meyers BC. 2013. Biogenesis and function of rice small RNAs from non-coding RNA precursors. *Curr Opin Plant Biol* 16:170–179.

Axtell MJ 2009. The Small RNAs of *Physcomitrella patens*: Expression, Function and Evolution. In Annual Plant Reviews Volume 36: The Moss *Physcomitrella patens* (eds C.D. Knight, P.-F. Perroud and D.J. Cove).

Axtell MJ. 2013. Classification and Comparison of Small RNAs from Plants. *Annu Rev Plant Biol* 64:137–159.

Axtell MJ, Jan C, Rajagopalan R, Bartel DP. 2006. A Two-Hit Trigger for siRNA Biogenesis in Plants. *Cell* 127:565–577.

Axtell MJ, Meyers BC. 2018. Revisiting criteria for plant miRNA annotation in the era of big data. *Plant Cell* 30:tpc.00851.2017.

Axtell MJ, Snyder JA, Bartel DP. 2007. Common Functions for Diverse Small RNAs of Land Plants. *Plant Cell Online* 19:1750–1769.

Bajczyk M, Lange H, Bielewicz D, Szewc L, Bhat SS, Dolata J, Kuhn L, Szweykowska-Kulinska Z, Gagliardi D, Jarmolowski A. 2020. SERRATE interacts with the nuclear exosome targeting (NEXT) complex to degrade primary miRNA precursors in Arabidopsis. *Nucleic Acids Res* 48:gkaa373.

Baldrich P, Bélanger S, Kong S, Pokhrel S, Tamim S, Teng C, Schiebout C, Gurazada SGR, Gupta P, Patel P, et al. 2021. The Evolutionary History of Small RNAs in the Solanaceae. *Biorxiv*:2021.05.26.445884.

Bélanger S, Pokhrel S, Czymmek KJ, Meyers BC. 2020. Pre-meiotic, 24-nt reproductive phasiRNAs are abundant in anthers of wheat and barley but not rice and maize. *Plant Physiol*:pp.00816.2020.

Borges F, Martienssen RA. 2015. The expanding world of small RNAs in plants. *Nat Rev Mol Cell Bio* 16:727–741.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–

1973.

Castro E de, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34:W362–W365.

Cho SH, Coruh C, Axtell MJ. 2012. miR156 and miR390 Regulate tasiRNA Accumulation and Developmental Timing in Physcomitrella patens. Plant Cell 24:4837–4849.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* 9:938–950.

Curtin SJ, Watson JM, Smith NA, Eamens AL, Blanchard CL, Waterhouse PM. 2008. The roles of plant dsRNA-binding proteins in RNAi-like pathways. *Febs Lett* 582:2753–2760.

Ding S-W, Voinnet O. 2007. Antiviral Immunity Directed by Small RNAs. Cell 130:413–426.

Duan C, Zhang H, Tang K, Zhu X, Qian W, Hou Y, Wang B, Lang Z, Zhao Y, Wang X, et al. 2015. Specific but interdependent functions for Arabidopsis AGO4 and AGO6 in RNA-directed DNA methylation. *Embo J* 34:581–592.

Dubcovsky J, Dvorak J. 2007. Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. *Science* 316:1862–1866.

Edwards KD, Fernandez-Pozo N, Drake-Stowe K, Humphry M, Evans AD, Bombarely A, Allen F, Hurst R, White B, Kernodle SP, et al. 2017. A reference genome for Nicotiana tabacum enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *Bmc Genomics* 18:448.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16:157.

Fei Q, Xia R, Meyers BC. 2013. Phased, Secondary, Small Interfering RNAs in Posttranscriptional Regulatory Networks. Plant Cell Online 25:2400–2415.

Fei Q, Yang L, Liang W, Zhang D, Meyers BC. 2016. Dynamic changes of small RNAs in rice spikelet development reveal specialized reproductive phasiRNA pathways. *J Exp Bot* 67:6037–6049.

Fukudome A, Fukuhara T. 2017. Plant dicer-like proteins: double-stranded RNA-cleaving enzymes for small RNA biogenesis. *J Plant Res* 130:33–44.

Garcia-Ruiz H, Takeda A, Chapman EJ, Sullivan CM, Fahlgren N, Brempelis KJ, Carrington JC. 2010. Arabidopsis RNA-Dependent RNA Polymerases and Dicer-

Like Proteins in Antiviral Defense and Small Interfering RNA Biogenesis during Turnip Mosaic Virus Infection. Plant Cell Online 22:481–496.

Gasciolli V, Mallory AC, Bartel DP, Vaucheret H. 2005. Partially Redundant Functions of Arabidopsis DICER-like Enzymes and a Role for DCL4 in Producing trans-Acting siRNAs. *Curr Biol* 15:1494–1500.

Gattiker A, Gasteiger E, Bairoch A. 2002. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinform* 1:107–108.

Hahn MW. 2009. Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *J Hered* 100:605–617.

Harvey JJW, Lewsey MG, Patel K, Westwood J, Heimstädt S, Carr JP, Baulcombe DC. 2011. An Antiviral Defense Role of AGO2 in Plants. *Plos One* 6:e14639.

Havecker ER, Wallbridge LM, Hardcastle TJ, Bush MS, Kelly KA, Dunn RM, Schwach F, Doonan JH, Baulcombe DC. 2010. The Arabidopsis RNA-Directed DNA Methylation Argonautes Functionally Diverge Based on Their Expression and Interaction with Target Loci. Plant Cell Online 22:321–334.

Herr AJ, Jensen MB, Dalmay T, Baulcombe DC. 2005. RNA Polymerase IV Directs Silencing of Endogenous DNA. Science 308:118–120.

Hiraguri A, Itoh R, Kondo N, Nomura Y, Aizawa D, Murai Y, Koiwa H, Seki M, Shinozaki K, Fukuhara T. 2005. Specific interactions between Dicer-like proteins and HYL1/DRB- family dsRNA-binding proteins in Arabidopsis thaliana. *Plant Mol Biol* 57:173–188.

Hua X, Berkowitz ND, Willmann MR, Yu X, Lyons E, Gregory BD. 2021. Global Analysis of RNA-Dependent RNA Polymerase-Dependent Small RNAs Reveals New Substrates and Functions for These Proteins and SGS3 in Arabidopsis. *Non-coding Rna* 7:28.

IWGSC. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:eaar7191.

Jha V, Narjala A, Basu D, N. ST, Pachamuthu K, Chenna S, Nair A, Shivaprasad PV. 2021. Essential role of γ-clade RNA-dependent RNA polymerases in rice development and yield-related traits is linked to their atypical polymerase activities regulating specific genomic regions. *New Phytol* 232:1674–1691.

Jia J, Ji R, Li Z, Yu Y, Nakano M, Long Y, Feng L, Qin C, Lu D, Zhan J, et al. 2020. Soybean Dicer-Like 2 Regulates Seed Coat Color via Production of Primary 22-nt Small Interfering RNAs from Long Inverted Repeats. Plant Cell 32:tpc.00562.2020.

Jiang P, Lian B, Liu C, Fu Z, Shen Y, Cheng Z, Qi Y. 2020. 21-nt phasiRNAs direct target mRNA cleavage in rice male germ cells. *Nat Commun* 11:5191.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, et al. 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biol 13:R3–R3.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473:97–100.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler A von, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589.

Kapoor M, Arora R, Lama T, Nijhawan A, Khurana JP, Tyagi AK, Kapoor S. 2008. Genome-wide identification, organization and phylogenetic analysis of Dicer-like, Argonaute and RNA-dependent RNA Polymerase gene families and their expression analysis during reproductive development and stress in rice. Bmc Genomics 9:451.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30:772–780.

Kirino Y, Mourelatos Z. 2007. The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. Rna 13:1397–1401.

Komiya R, Ohyanagi H, Niihama M, Watanabe T, Nakano M, Kurata N, Nonomura K. 2014. Rice germline-specific Argonaute MEL1 protein binds to phasiRNAs generated from more than 700 lincRNAs. Plant J 78:385–397.

Lee Y-S, Maple R, Dürr J, Dawson A, Tamim S, Genio C del, Papareddy R, Luo A, Lamb JC, Amantia S, et al. 2021. A transposon surveillance mechanism that safeguards plant male fertility during stress. *Nat Plants* 7:34–41.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:gkz239-.

Li Y, Huang Y, Pan L, Zhao Y, Huang W, Jin W. 2021. Male sterile 28 encodes an ARGONAUTE family protein essential for male fertility in maize. *Chromosome Res*:1–13.

Li Z, Li W, Guo M, Liu S, Liu L, Yu Y, Mo B, Chen X, Gao L. 2021. Origin, evolution and diversification of plant ARGONAUTE proteins. Plant J. *In Press*

Liu B, Chen Z, Song X, Liu C, Cui X, Zhao X, Fang J, Xu W, Zhang H, Wang X, et al. 2007. Oryza sativa Dicer-like4 Reveals a Key Role for Small Interfering RNA Silencing in Plant Development. Plant Cell 19:2705–2718.

Liu Y, Teng C, Xia R, Meyers BC. 2020. PhasiRNAs in Plants: Their Biogenesis, Genic Sources, and Roles in Stress Responses, Development, and Reproduction. Plant Cell 32:3059–3080.

Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade VM, Milner SG, et al. 2019. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet* 51:885–895.

MacRae IJ, Zhou K, Doudna JA. 2007. Structural determinants of RNA recognition and cleavage by Dicer. *Nat Struct Mol Biol* 14:934–940.

MacRae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, Adams PD, Doudna JA. 2006. Structural Basis for Double-Stranded RNA Processing by Dicer. Science 311:195–198.

Mallory A, Vaucheret H. 2010. Form, Function, and Regulation of ARGONAUTE Proteins. Plant Cell 22:3879–3889.

Minh BQ, Nguyen MAT, Haeseler A von. 2013. Ultrafast Approximation for Phylogenetic Bootstrap. Mol Biol Evol 30:1188–1195.

Montgomery TA, Howell MD, Cuperus JT, Li D, Hansen JE, Alexander AL, Chapman EJ, Fahlgren N, Allen E, Carrington JC. 2008. Specificity of ARGONAUTE7-miR390 Interaction and Dual Functionality in TAS3 Trans-Acting siRNA Formation. Cell 133:128–141.

Nakazawa Y, Hiraguri A, Moriyama H, Fukuhara T. 2007. The dsRNA-binding protein DRB4 interacts with the Dicer-like protein DCL4 in vivo and functions in the trans-acting siRNA pathway. Plant Mol Biol 63:777–785.

Nguyen L-T, Schmidt HA, Haeseler A von, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 32:268–274.

Nonomura K-I, Morohoshi A, Nakano M, Eiguchi M, Miyao A, Hirochika H, Kurata N. 2007. A Germ Cell–Specific Gene of the ARGONAUTE Family Is Essential for the Progression of Premeiotic Mitosis and Meiosis during Sporogenesis in Rice. Plant Cell 19:2583–2594.

Olmedo-Monfil V, Durán-Figueroa N, Arteaga-Vázquez M, Demesa-Arévalo E, Autran D, Grimanelli D, Slotkin RK, Martienssen RA, Vielle-Calzada J-P. 2010. Control of female gamete formation by a small RNA pathway in Arabidopsis. Nature 464:628–632.

Parent J, Bouteiller N, Elmayan T, Vaucheret H. 2015. Respective contributions of Arabidopsis DCL2 and DCL4 to RNA silencing. Plant J 81:223–232.

Patel P, Mathioni SM, Hammond R, Harkess AE, Kakrana A, Arikit S, Dusia A, Meyers

BC. 2021. Reproductive phasiRNA loci and DICER-LIKE5, but not microRNA loci, diversified in monocotyledonous plants. Plant Physiol:kiab001.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *P Natl Acad Sci Usa* 101:9903–9908.

Peng Yuanying, Yan H, Guo L, Deng C, Kang L, Wang C, Zhou P, Yu K, Dong X, Zhao J, et al. 2021. Reference genome assemblies reveal the origin and evolution of allohexaploid oat.

Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS. 2004. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Gene Dev* 18:2368–2379.

Pokhrel S, Huang K, Bélanger S, Caplan JL, Kramer EM, Meyers BC. 2020. Pre-meiotic, 21-nucleotide Reproductive PhasiRNAs Emerged in Seed Plants and Diversified in Flowering Plants. Biorxiv:2020.10.16.341925.

Pokhrel S, Huang K, Bélanger S, Zhan J, Caplan JL, Kramer EM, Meyers BC. 2021. Pre-meiotic 21-nucleotide reproductive phasiRNAs emerged in seed plants and diversified in flowering plants. Nat Commun 12:4941.

Pozo JC del, Ramirez-Parra E. 2015. Whole genome duplications in plants: an overview from Arabidopsis. J Exp Bot 66:6991–7003.

Rowley MJ, Avrutsky MI, Sifuentes CJ, Pereira L, Wierzbicki AT. 2011. Independent Chromatin Binding of ARGONAUTE4 and SPT5L/KTF1 Mediates Transcriptional Gene Silencing. *Plos Genet* 7:e1002120.

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C. 2008. Identification and Characterization of Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution. Plant Cell 20:11–24.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr Opin Plant Biol 15:147–153.

Singh M, Goel S, Meeley RB, Dantec C, Parrinello H, Michaud C, Leblanc O, Grimanelli D. 2011. Production of Viable Gametes without Meiosis in Maize Deficient for an ARGONAUTE Protein. Plant Cell 23:443–458.

Song X, Li P, Zhai J, Zhou M, Ma L, Liu B, Jeong D, Nakano M, Cao S, Liu C, et al. 2012. Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J 69:462–474.

Sun W, Chen D, Xue Y, Zhai L, Zhang D, Cao Z, Liu L, Cheng C, Zhang Y, Zhang Z. 2019. Genome-wide identification of AGO18b-bound miRNAs and phasiRNAs in

maize by cRIP-seq. Bmc Genomics 20:656.

Sun W, Xiang X, Zhai L, Zhang D, Cao Z, Liu L, Zhang Z. 2018. AGO18b negatively regulates determinacy of spikelet meristems on the tassel central spike in maize. *J Integr Plant Biol* 60:65–78.

Talmor-Neiman M, Stav R, Klipcan L, Buxdorf K, Baulcombe DC, Arazi T. 2006. Identification of trans-acting siRNAs in moss and an RNA-dependent RNA polymerase required for their biogenesis. Plant J 48:511–521.

Tamim S, Cai Z, Mathioni SM, Zhai J, Teng C, Zhang Q, Meyers BC. 2018. Cis-directed cleavage and nonstoichiometric abundances of 21-nucleotide reproductive phased small interfering RNAs in grasses. New Phytol 220:865–877.

Taochy C, Gursanscky NR, Cao J, Fletcher SJ, Dressel U, Mitter N, Tucker MR, Koltunow AMG, Bowman JL, Vaucheret H, et al. 2017. A Genetic Screen for Impaired Systemic RNAi Highlights the Crucial Role of DICER-LIKE 2. Plant Physiol 175:1424–1437.

Teng C, Zhang H, Hammond R, Huang K, Meyers BC, Walbot V. 2020. Dicer-like 5 deficiency confers temperature-sensitive male sterility in maize. Nat Commun 11:2912.

Vaucheret H. 2008. Plant ARGONAUTES. *Trends Plant Sci* 13:350–358.

Wang Q, Xue Y, Zhang L, Zhong Z, Feng S, Wang C, Xiao L, Yang Z, Harris CJ, Wu Z, et al. 2021. Mechanism of siRNA production by a plant Dicer-RNA complex in dicing-competent conformation. Science 374:1152–1157.

Wang S, Liang H, Xu Y, Li L, Wang H, Sahu DN, Petersen M, Melkonian M, Sahu SK, Liu H. 2021. Genome-wide analyses across Viridiplantae reveal the origin and diversification of small RNA pathway-related genes. *Commun Biology* 4:412.

Wang Z, Hardcastle TJ, Pastor AC, Yip WH, Tang S, Baulcombe DC. 2018. A novel DCL2-dependent miRNA pathway in tomato affects susceptibility to RNA viruses. Gene Dev 32:1155–1160.

Willmann MR, Endres MW, Cook RT, Gregory BD. 2011. The Functions of RNA-Dependent RNA Polymerases in Arabidopsis. *Arabidopsis Book* 2011:e0146.

Wu J, Yang Z, Wang Y, Zheng L, Ye R, Ji Y, Zhao S, Ji S, Liu R, Xu L, et al. 2015. Viral-inducible Argonaute18 confers broad-spectrum virus resistance in rice by sequestering a host microRNA. *Elife* 4:e05733.

Wu L, Mao L, Qi Y. 2012. Roles of DICER-LIKE and ARGONAUTE Proteins in TAS-Derived Small Interfering RNA-Triggered DNA Methylation. Plant Physiol 160:990–999.

Wu Y, Hou B, Lee W, Lu S, Yang C, Vaucheret H, Chen H. 2017. DCL2- and RDR6-

dependent transitive silencing of SMXL4 and SMXL5 in Arabidopsis dcl4 mutants causes defective phloem transport and carbohydrate over-accumulation. Plant J 90:1064–1078.

Xia R, Chen C, Pokhrel S, Ma W, Huang K, Patel P, Wang F, Xu J, Liu Z, Li J, et al. 2019. 24-nt reproductive phasiRNAs are broadly present in angiosperms. Nat Commun 10:627.

Xia R, Xu J, Meyers BC. 2017. The Emergence, Evolution, and Diversification of the miR390-TAS3-ARF Pathway in Land Plants. Plant Cell 29:1232–1247.

Xie Z, Allen E, Wilken A, Carrington JC. 2005. DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in Arabidopsis thaliana. P Natl Acad Sci Usa 102:12984–12989.

Yang L, Liu Z, Lu F, Dong A, Huang H. 2006. SERRATE is a novel nuclear regulator in primary microRNA processing in Arabidopsis. Plant J 47:841–850.

Yang Z, Ebright YW, Yu B, Chen X. 2006. HEN1 recognizes 21-24 nt small RNA duplexes and deposits a methyl group onto the 2’ OH of the 3’ terminal nucleotide. Nucleic Acids Res 34:667–675.

You C, Cui J, Wang H, Qi X, Kuo L-Y, Ma H, Gao L, Mo B, Chen X. 2017. Conservation and divergence of small RNA pathways and microRNAs in land plants. Genome Biol 18:158.

Zhai J, Zhang H, Arikit S, Huang K, Nan G-L, Walbot V, Meyers BC. 2015. Spatiotemporally dynamic, cell-type–dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proc National Acad Sci* 112:3146–3151.

Zhang H, Xia R, Meyers BC, Walbot V. 2015. Evolution, functions, and mysteries of plant ARGONAUTE proteins. Curr Opin Plant Biol 27:84–90.

Zhang M, Ma X, Wang C, Li Q, Meyers BC, Springer NM, Walbot V. 2021. CHH DNA methylation increases at 24-PHAS loci depend on 24-nt phased small interfering RNAs in maize meiotic anthers. New Phytol 229:2984–2997.

Zhang Y-C, Lei M-Q, Zhou Y-F, Yang Y-W, Lian J-P, Yu Y, Feng Y-Z, Zhou K-R, He R-R, He H, et al. 2020. Reproductive phasiRNAs regulate reprogramming of gene expression and meiotic progression in rice. Nat Commun 11:6031.

Zheng X, Zhu Jianhua, Kapoor A, Zhu Jian-Kang. 2007. Role of Arabidopsis AGO6 in siRNA accumulation, DNA methylation and transcriptional gene silencing. Embo J 26:1691–1701.

FIGURES AND TABLES

Figure 1. Maximum likelihood tree of 1,375 RDR proteins identified in 198 Viridiplantae genomes and 21 outgroup species.

The RDRs form three major clades: RDR1/2, RDR3 and RDR6. The phylogenetic tree was rooted at the branch that separates outgroup and green algae species to land plants. Bootstrap values (1000 replicates) less than 100 are marked, meaning that the absence of a bootstrap value indicates 100. Branches and leaves are color-coded representing following taxonomic groups: outgroup (gray), chlorophyte (pink), bryophyte (orange), lycophyte and fern (purple), gymnosperm (green), monocot (red) and eudicot (blue) in angiosperm. The species codes and their corresponding full species names are listed in Supplemental Table 1, and protein names in Supplemental Table 2. The annotated tree in Newick format is provided in Supplemental data 1.

Figure 2. Eudicots lost a copy of the DCL3 gene.

(A) Maximum likelihood tree of 1,098 DCL proteins found in 198 Viridiplantae genomes and 21 outgroup species. Rooted at the branch that separates metazoan, fungi and algae species to land plant species, DCL proteins form four major clades: DCL1, DCL2, DCL4 and DCL3/DCL5. (B) Pruned at the DCL3/DCL5 clade, the tree shows that the monocots gained a copy of DCL3 that diverged, relative to the eudicots; eudicot genomes encode a DCL3 protein that is sister to the monocot DCL5. Color coding, species codes and names, and bootstrap values are as described in Figure 1. The annotated tree in Newick format is provided in Supplemental Data 2.

Figure 3. Sequence divergence in specific domains underlies the functional divergence of DCL3, DCL4 and DCL5 proteins

(A) Rice DCL5 protein structure predicted by AlphaFold. (B) Conserved amino acids (identity > 95%) are colored in green when conserved among DCL3, DCL4 and *DCL5* (86 amino acids), black between DCL4 and *DCL5* (35 amino acids) and yellow black between DCL3 and DCL5 (46 amino acids). (C) Regions of amino acids conserved between DCL3 and DCL5 but absent in DCL4 are shown in red when located directly at

two junctions between the PAZ domain and adjacent domains or in olive when positioned on folded regions near by the PAZ domain.

Figure 4. Evolution of AGO proteins in clades AGO4/6 and AGO2/7.

AGO4/6 (A) and AGO2/7 (B) clades pruned from a Maximum likelihood tree of 2,991 AGO proteins. Trees show the plant lineages in which AGOs of these clades diverged. Colored panels highlight when these AGOs have been lost or retained after duplication events in specific lineages. Color coding, species codes and names, and bootstrap values are as described in Figure 1. The annotated tree in Newick format is provided in Supplemental Data 3.

Figure 5. AGO proteins in the AGO1/5/10/18 clade.

(A) Pruned tree of AGO1/5/10/18 shows the presence and absence of lineages in each group. The tree shows the plant lineages in which AGOs of these clades diverged. (B) AGO18 tree showing the presence of this AGO in the Cyperaceae and almost every Poaceae species but absent in other monocots. These clades were pruned from a Maximum likelihood tree of 2,991 AGO proteins. Color coding, species codes and names, and bootstrap values are as described in Figure 1. The annotated tree in Newick format is provided in Supplemental Data 3.

Figure 6. AGO proteins in the AGO1 and AGO5 clades.

AGO1 (A) and AGO5 (B) clades pruned from a Maximum likelihood tree of 2,991 AGO proteins. Trees show the plant lineages in which AGOs of these clades diverged. We highlight the absence (shaded blue panel) and presence (shaded green panel) of AGOs specific to monocot subgroups. Color coding, species codes and names, and bootstrap values are as described in Figure 1. The annotated tree in Newick format is provided in Supplemental Data 3.

Table 1. Features of key AGO, DCL and RDR proteins known to function in the hc-siRNA and phasiRNA pathways.

Protein	Plant model	Features		References
		sRNA length (nt)[1]	siRNA pathway	
DCL3	Arabidopsis	24	hc-siRNA	(Q. Wang et al. 2021)
DCL4	Arabidopsis, rice	21	phasiRNA	(Xie et al. 2005; Liu et al. 2007; Song et al. 2012)
DCL5	Maize, rice	24	phasiRNA	(Song et al. 2012; Teng et al. 2020)
RDR2	Arabidopsis	24	hc-siRNA	(Willmann et al. 2011; Hua et al. 2021)
RDR6	Arabidopsis, rice	24	phasiRNA	(Peragine et al. 2004; Willmann et al. 2011; Hua et al. 2021)
AGO5c	Rice, maize	21	phasiRNA	(Komiya et al. 2014; Lee et al. 2021)
AGO4	Arabidopsis	24	hc-siRNA	(Rowley et al. 2011)
AGO6	Arabidopsis	24	hc-siRNA	(Zheng et al. 2007)
AGO18	Maize, rice	21 and 24	phasiRNA	(Sun et al. 2019)

[1]Typical length of sRNAs associated with these proteins.

Table 2. Comparison of the numbers of AGO proteins annotated in this study versus the Li et al. (2021) study.

Taxonomy	This study			Study from Li et al. 2021		
	Number of species	Number of AGOs	Average number of AGOs	Number of species	Number of AGOs	Average number of AGOs
Rhodophyta	5	8	1.6	-	-	-
Chlorophyta	11	13	1.2	3	9	3.0
Charophyta	-	-	-	13	48	3.7
Non-seed land plants						
Marchantiophyta and early lineage	2	6	3.0	-	-	-
Bryophyta	4	37	9.3	5	29	5.8
Lycopodiopsida	1	5	5.0	1	7	7.0
Polypodiopsida	2	14	7.0	2	12	6.0
Acrogymnospermae	8	74	9.3	5	34	6.8
Mesangiospermae	170	2,746	16.2	215	2,819	13.1
Basal angiosperm	3	31	10.3	-	-	-
Liliopsida	48	986	20.5	-	-	-
Pentapetalae	119	1,729	14.5	-	-	-

SUPPLEMENTAL FIGURES AND DATA

Supplemental Figure 1. Functional domains conserved in the seven protein families studied in this work.

Each domain is color-coded. The position and size of domains are proportional within and between proteins. Arabidopsis proteins were used as models.

Supplemental Figure 2. Phylogeny of plant species analyzed in this study.

The species tree was generated using the NCBI common taxonomy tree tool for 219 species, including 198 Viridiplantae, and visualized using iTOL. Colored panels highlight specific lineages.

Supplemental Figure 3. Alignment of rice DCL5 protein structure predicted by AlphaFold to the DCL3 structure experimentally described in Arabidopsis.

Known protein domains were color-coded in the rice DCL5 structures predicted by AlphaFold. The Arabidopsis DCL3*,* experimentally elucidated, is in black, except the sRNA molecule. The alignment shows that *alpha* helices and *beta* sheets predicted with AlphaFold overlap well with the structure experimentally elucidated in Arabidopsis

Supplemental Figure 4. Maximum likelihood tree of 2,991 AGO proteins.

The unrooted tree shows the presence of eight AGO clades present in Viridiplantae, namely AGO6, AGO4, AGO7, AGO2, AGO18, AGO5, AGO10, and AGO1. Bootstrap values (1000 replicates) less than 100 are marked, meaning that the absence of a bootstrap value indicates 100. Branches and leaves are color-coded representing following taxonomic groups: outgroup (gray), chlorophyte (pink), bryophyte (orange), lycophyte and fern (purple), gymnosperm (green), monocot (red) and eudicot (blue) in angiosperm. The species codes and their corresponding full species names are listed in Supplemental Table 1, and protein names in Supplemental Table 2. The annotated tree in Newick format is provided in Supplemental Data 3.

Supplemental Figure 5. Maximum likelihood tree of 1,903 DRB proteins.

The topology shows that DRBs group in five clades: DRB1 to DRB5. The tree was rooted at the branch that separates outgroup and algae species to land plants. Color coding, species codes and names, and bootstrap values are as described in Supplemental Figure 4. The annotated tree in Newick format is provided in Supplemental Data 4.

Supplemental Figure 6. Maximum likelihood tree of 286 HEN1 proteins.

HEN1 proteins form one group that mimics the species tree. Yellow panels indicate a duplication of HEN1 in Brassicaceae. Arabidopsis and rice reference genes are in bold and highlighted in gray. The tree was rooted at the branch that separates the outgroup to green algaes and land plants. Color coding, species codes and names, and bootstrap values are as described in Supplemental Figure 4.The annotated tree in Newick format is provided in Supplemental Data 5.

Supplemental Figure 7. Maximum likelihood tree of 486 SE proteins.

SE proteins form one group. Yellow panels mark duplicated groups of SE observed in Poales. Arabidopsis and rice reference genes are in bold and highlighted in gray. The tree was rooted at the branch that separates the outgroup to algaes and land plants. Color coding, species codes and names, and bootstrap values are as described in Supplemental Figure 4. The annotated tree in Newick format is provided in Supplemental data 6.

Supplemental Figure 8. Maximum likelihood tree of 465 SGS3 proteins.

No SGS3 proteins were detected from outgroup species and green algae. The tree was rooted at *Marchantia polymorpha*. Color coding, species codes and names, and bootstrap values are as described in Supplemental Figure 4.The annotated tree in Newick format is provided in Supplemental Data 7.

Supplemental Data 1. Maximum likelihood tree of 1,375 RDR proteins in Newick format.

Supplemental Data 2. Maximum likelihood tree of 1,098 DCL proteins in Newick format.

Supplemental Data 3. Maximum likelihood tree of 2,991 AGO proteins in Newick format.

Supplemental Data 4. Maximum likelihood tree of 1,903 DRB proteins in Newick format.

Supplemental Data 5. Maximum likelihood tree of 286 HEN1 proteins in Newick format.

Supplemental Data 6. Maximum likelihood tree of 486 SE proteins in Newick format.

Supplemental Data 7. Maximum likelihood tree of 465 SGS3 proteins in Newick format.

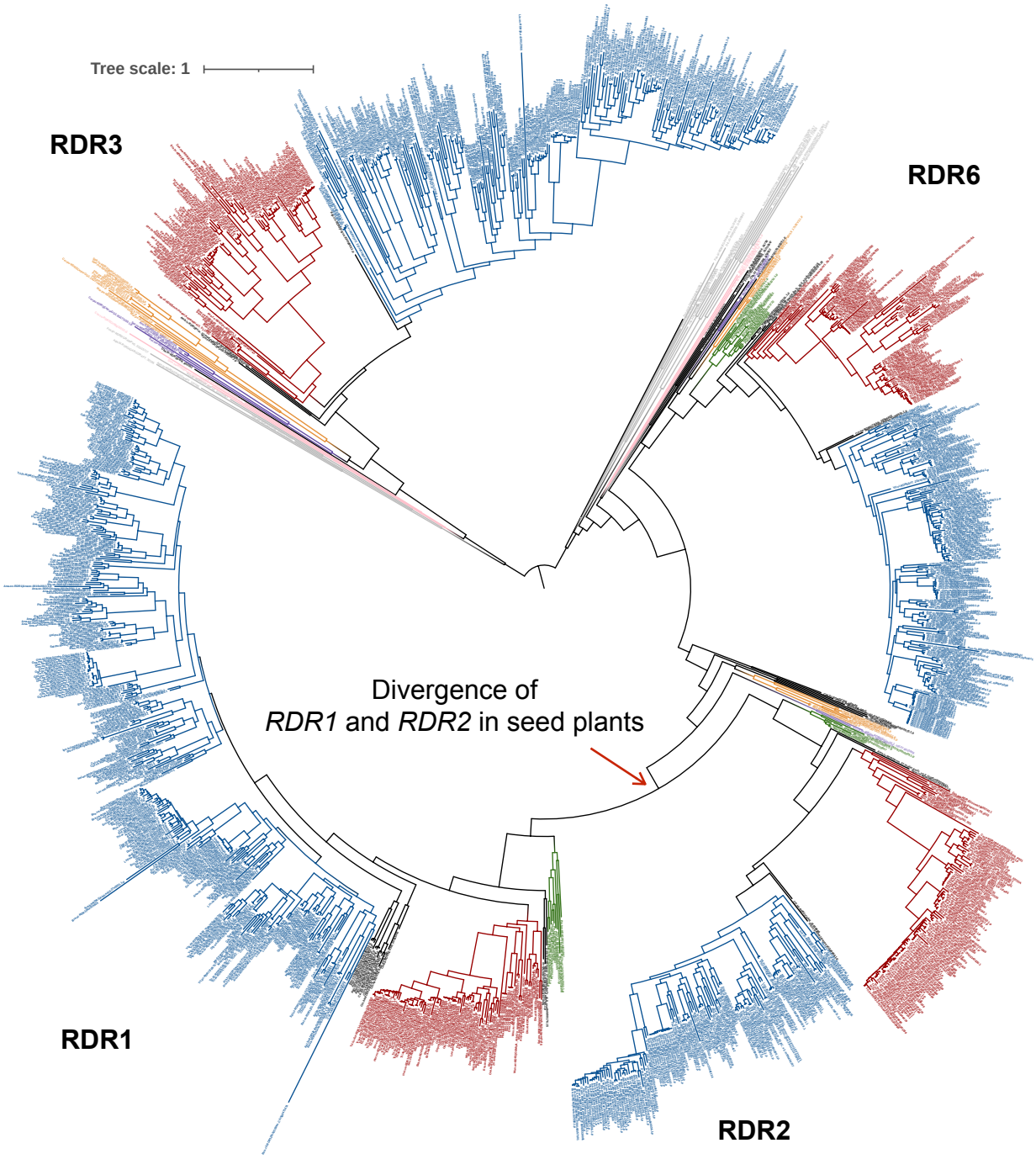

Tree scale: 1
RDR3
RDR6
Divergence of
RDR1 and *RDR2* in seed plants
RDR1
RDR2

Figure 1. Maximum likelihood tree of 1,375 RDR proteins identified in 198 Viridiplantae genomes and 21 outgroup species.

The RDRs form three major clades: RDR1/2, RDR3 and RDR6. The phylogenetic tree was rooted at the branch that separates outgroup and green algae species to land plants. Bootstrap values (1000 replicates) less than 100 are marked, meaning that the absence of a bootstrap value indicates 100. Branches and leaves are color-coded representing following taxonomic groups: outgroup (gray), chlorophyte (pink), bryophyte (orange), lycophyte and fern (purple), gymnosperm (green), monocot (red) and eudicot (blue) in angiosperm. The species codes and their corresponding full species names are listed in Supplemental Table 1, and protein names in Supplemental Table 2. The annotated tree in Newick format is provided in Supplemental data 1.

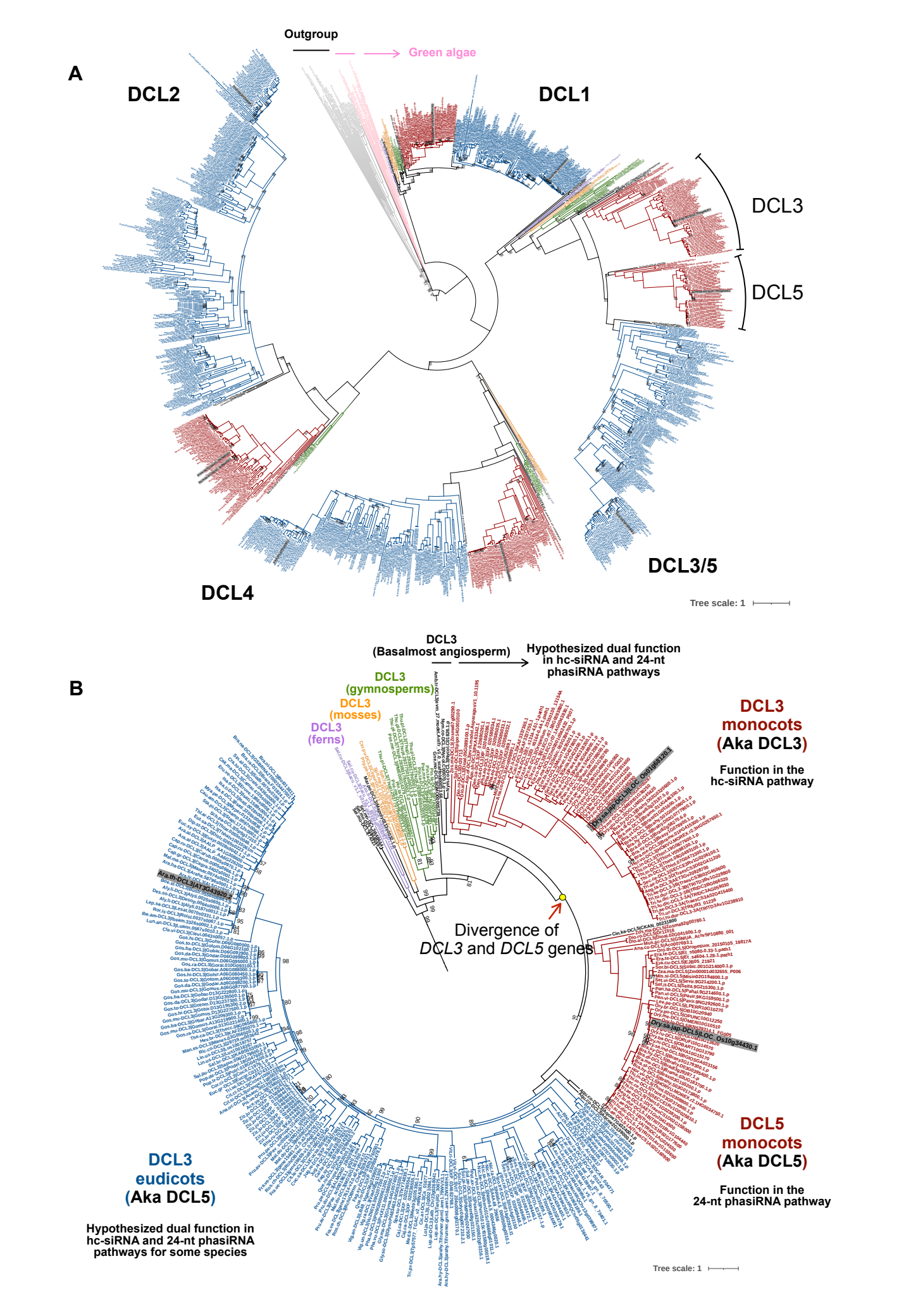

A
Outgroup
Green algae
DCL2
DCL1
DCL3
DCL5
DCL3/5
DCL4
Tree scale: 1
B
DCL3
(Basalmost angiosperm)
Hypothesized dual function
in hc-siRNA and 24-nt
phasiRNA pathways
DCL3
(gymnosperms)
DCL3
(mosses)
DCL3
(ferns)
DCL3
monocots
(Aka DCL3)
Function in the
hc-siRNA pathway
Divergence of
DCL3 and DCL5 genes
DCL5
monocots
(Aka DCL5)
Function in the
24-nt phasiRNA pathway
DCL3
eudicots
(Aka DCL5)
Hypothesized dual function in
hc-siRNA and 24-nt phasiRNA
pathways for some species
Tree scale: 1

Figure 2. Eudicots lost a copy of the DCL3 gene.

(A) Maximum likelihood tree of 1,098 DCL proteins found in 198 Viridiplantae genomes and 21 outgroup species. Rooted at the branch that separates metazoan, fungi and algae species to land plant species, DCL proteins form four major clades: DCL1, DCL2, DCL4 and DCL3/DCL5. (B) Pruned at the DCL3/DCL5 clade, the tree shows that the monocots gained a copy of DCL3 that diverged, relative to the eudicots; eudicot genomes encode a DCL3 protein that is sister to the monocot DCL5. Color coding, species codes and names, and bootstrap values are as described in Figure 1. The annotated tree in Newick format is provided in Supplemental Data 2.

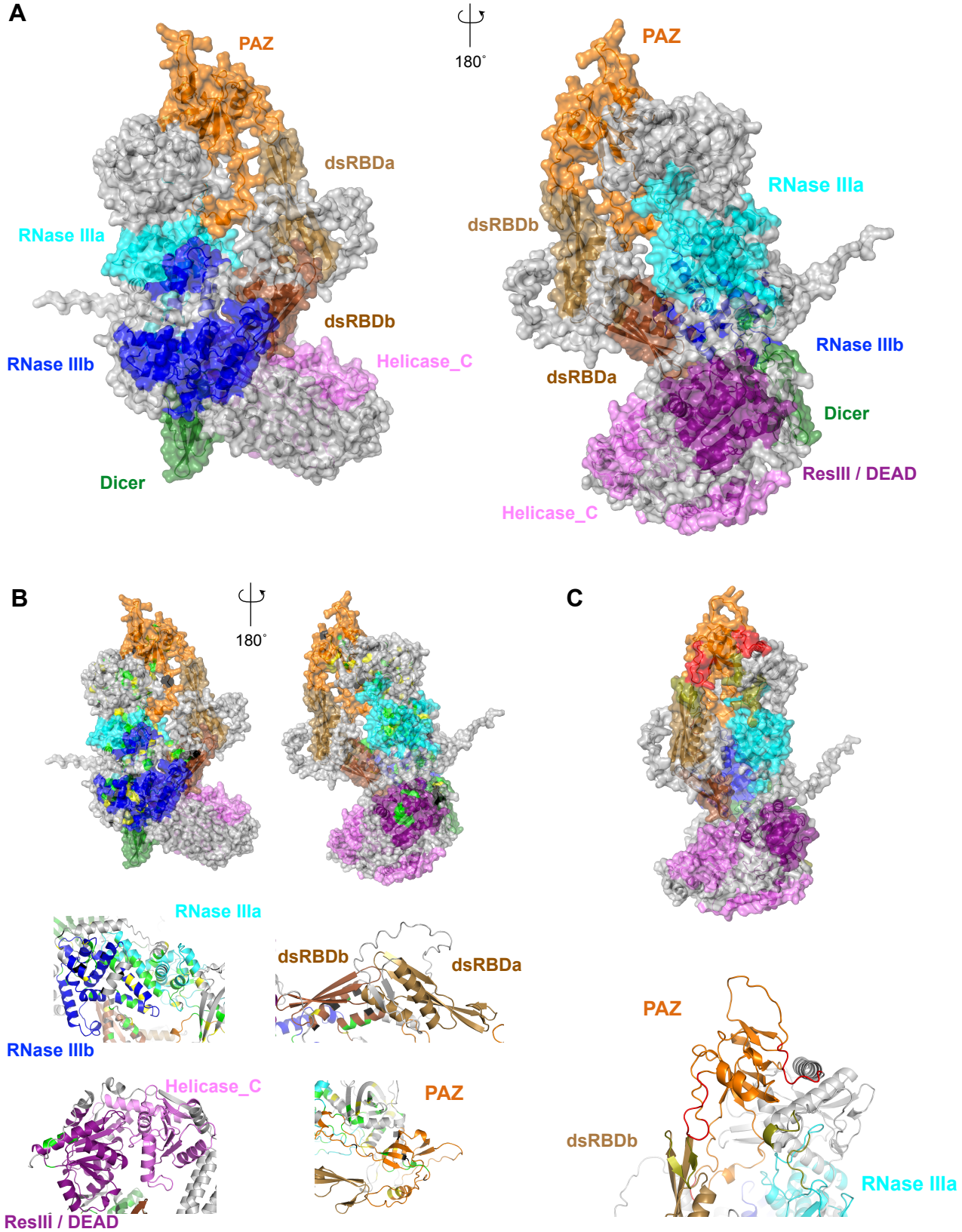
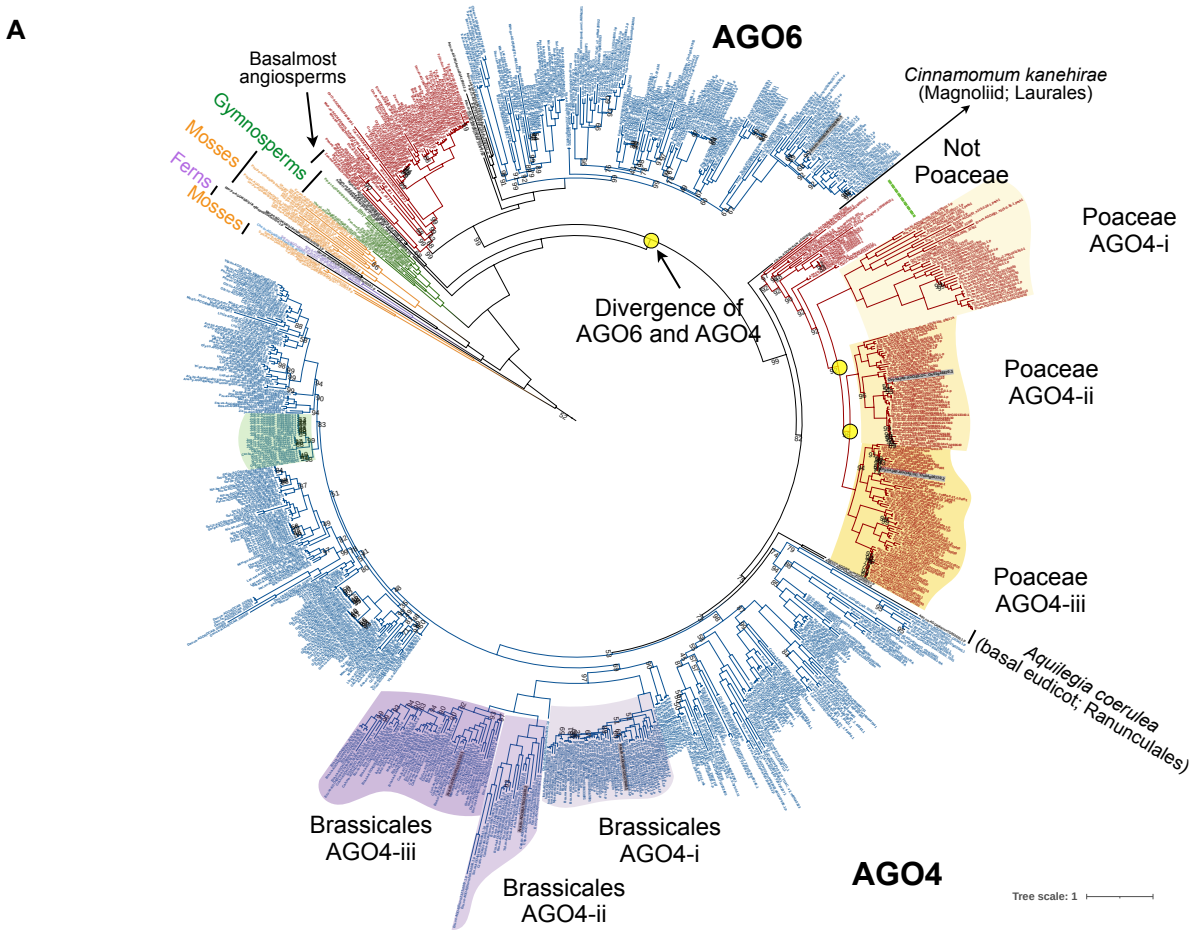

A
PAZ
dsRBDa
RNase IIIa
dsRBDb
RNase IIIb
Helicase_C
Dicer
180˚
PAZ
RNase IIIa
dsRBDb
RNase IIIb
dsRBDa
Dicer
ResIII / DEAD
Helicase_C
B
180˚
RNase IIIa
dsRBDb
dsRBDa
RNase IIIb
Helicase_C
PAZ
ResIII / DEAD
C
PAZ
dsRBDb
RNase IIIa

Figure 3. Sequence divergence in specific domains underlies the functional divergence of DCL3, DCL4 and DCL5 proteins

(A) Rice DCL5 protein structure predicted by AlphaFold. (B) Conserved amino acids (identity > 95%) are colored in green when conserved among DCL3, DCL4 and *DCL5* (86 amino acids), black between DCL4 and *DCL5* (35 amino acids) and yellow black between DCL3 and DCL5 (46 amino acids). (C) Regions of amino acids conserved between DCL3 and DCL5 but absent in DCL4 are shown in red when located directly at two junctions between the PAZ domain and adjacent domains or in olive when positioned on folded regions near by the PAZ domain.


A
AGO6
Basalmost angiosperms
Gymnosperms
Mosses
Ferns
Mosses
Cinnamomum kanehirae (Magnoliid; Laurales)
Not Poaceae
Poaceae AGO4-i
Divergence of AGO6 and AGO4
Poaceae AGO4-ii
Poaceae AGO4-iii
Aquilegia coerulea (basal eudicot; Ranunculales)
Brassicales AGO4-iii
Brassicales AGO4-i
Brassicales AGO4-ii
AGO4
Tree scale: 1

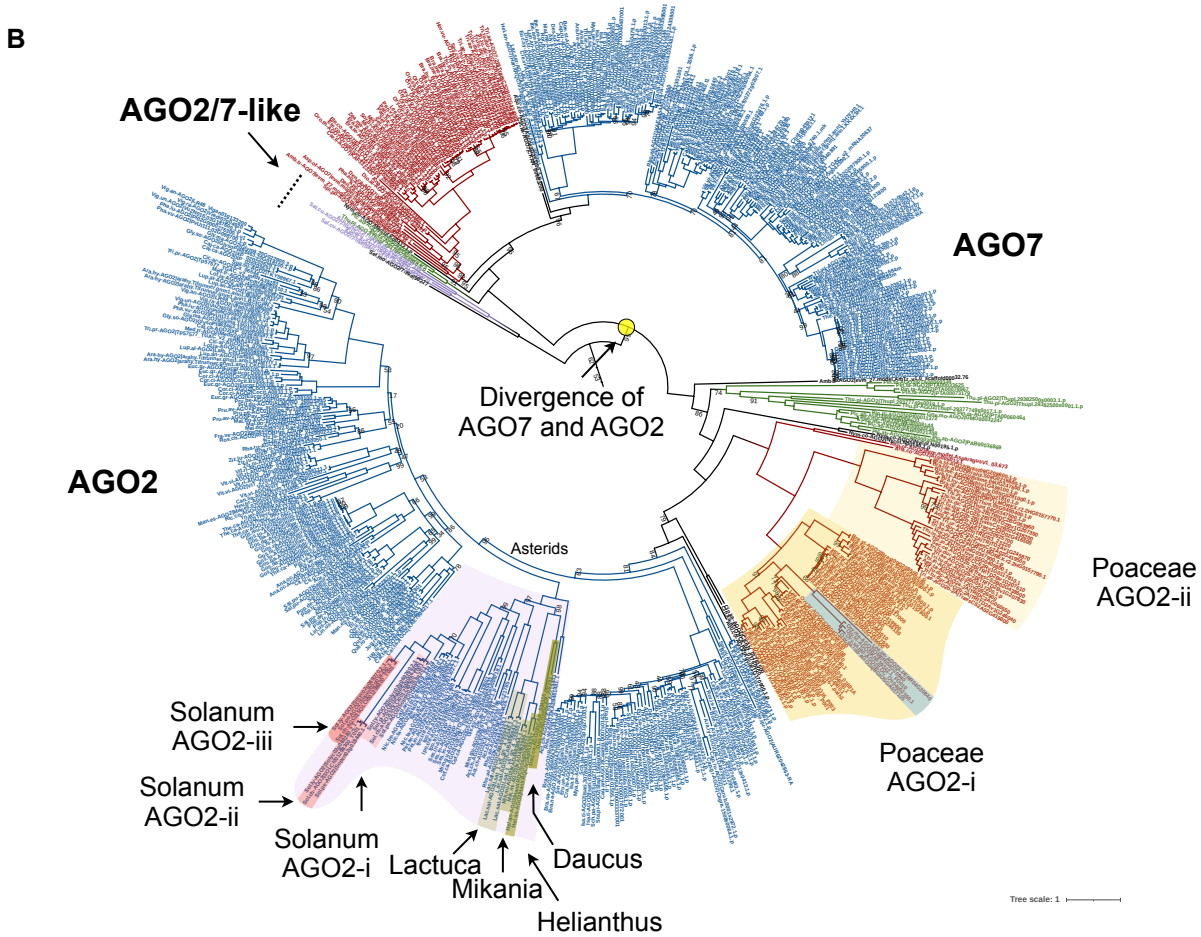
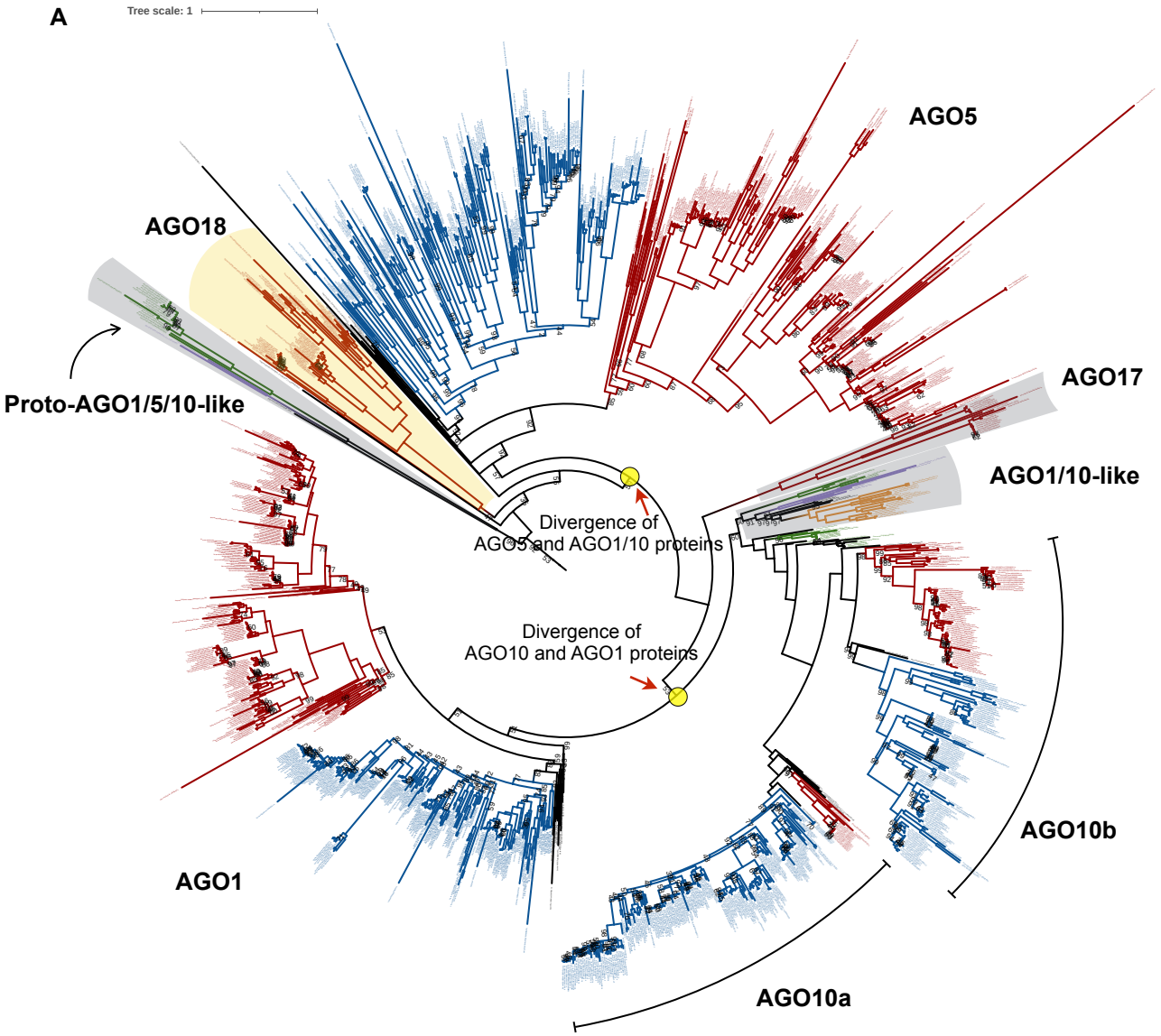

B
AGO2/7-like
AGO7
Divergence of AGO7 and AGO2
AGO2
Asterids
Poaceae AGO2-ii
Poaceae AGO2-i
Solanum AGO2-iii
Solanum AGO2-ii
Solanum AGO2-i
Lactuca
Mikania
Daucus
Helianthus
Tree scale: 1

Figure 4. Evolution of AGO proteins in clades AGO4/6 and AGO2/7.

AGO4/6 (A) and AGO2/7 (B) clades pruned from a Maximum likelihood tree of 2,991 AGO proteins. Trees show the plant lineages in which AGOs of these clades diverged. Colored panels highlight when these AGOs have been lost or retained after duplication events in specific lineages. Color coding, species codes and names, and bootstrap values are as described in Figure 1. The annotated tree in Newick format is provided in Supplemental Data 3.


A
Tree scale: 1
AGO5
AGO18
AGO17
Proto-AGO1/5/10-like
AGO1/10-like
Divergence of
AGO5 and AGO1/10 proteins
Divergence of
AGO10 and AGO1 proteins
AGO10b
AGO1
AGO10a

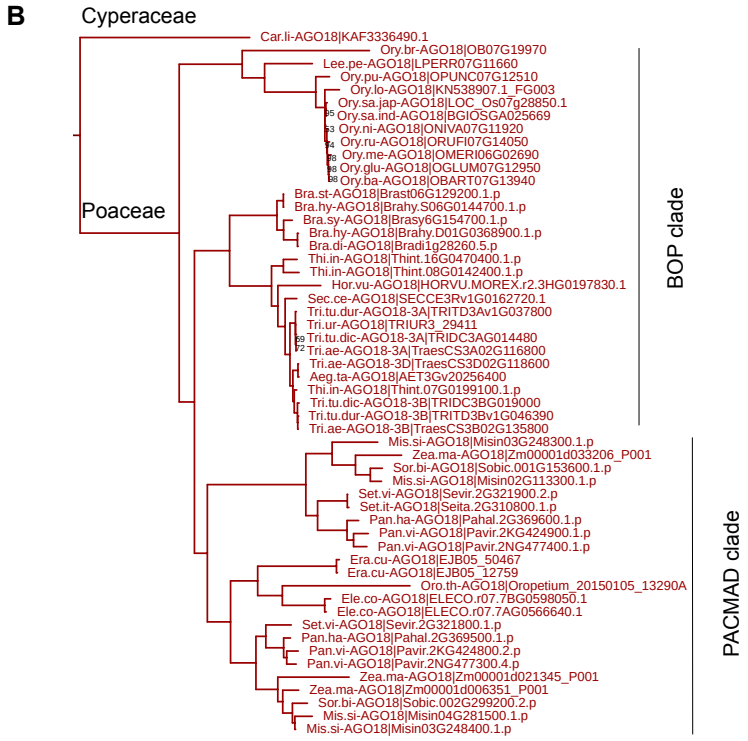
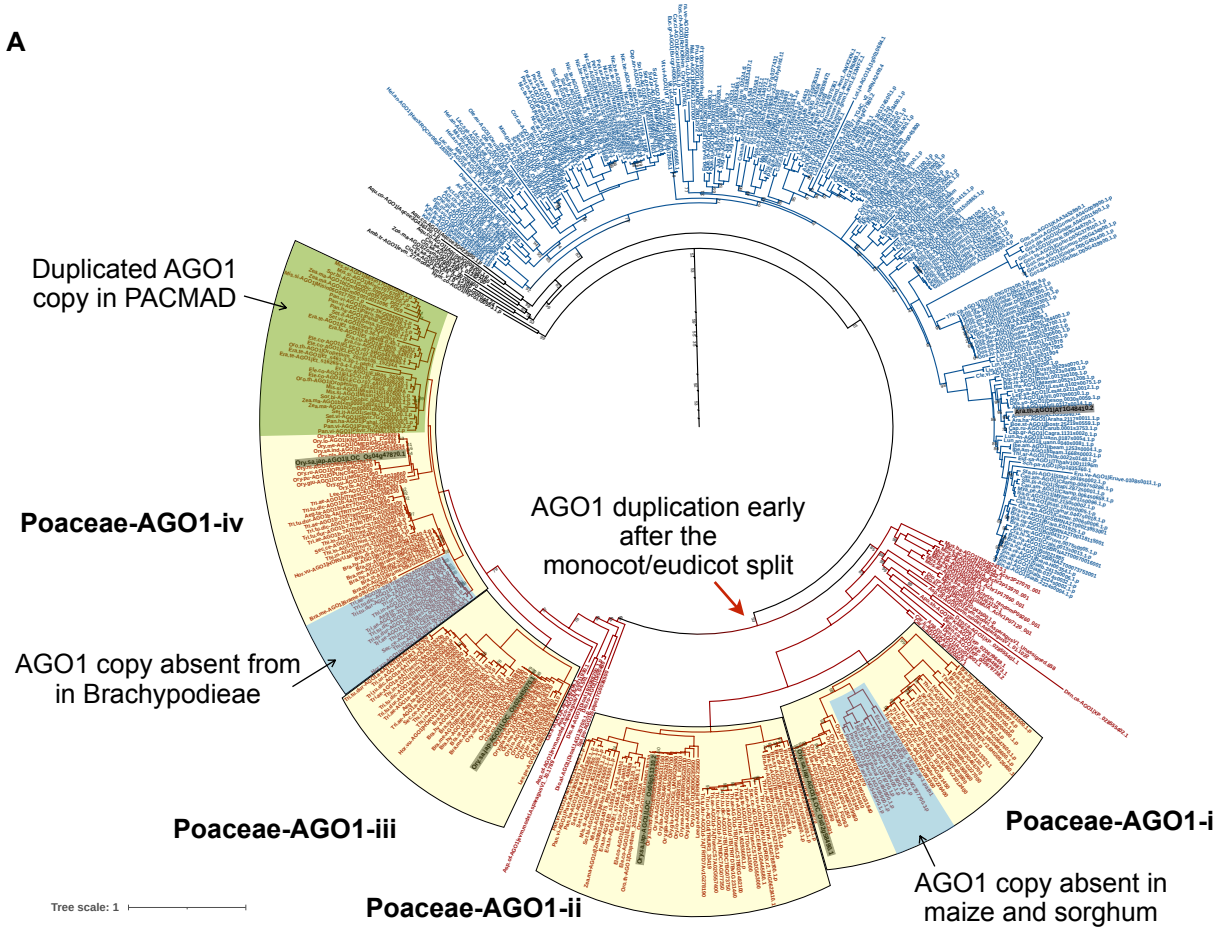

B
Cyperaceae
Poaceae
Car.li-AGO18|KAF3336490.1
Ory.br-AGO18|OB07G19970
Lee.pe-AGO18|LPERR07G11660
Ory.pu-AGO18|OPUNC07G12510
Ory.lo-AGO18|KN538907.1_FG003
Ory.sa.jap-AGO18|LOC_Os07g28850.1
Ory.sa.ind-AGO18|BGIOSGA025669
Ory.ni-AGO18|ONIVA07G11920
Ory.ru-AGO18|ORUFI07G14050
Ory.me-AGO18|OMERI06G02690
Ory.glu-AGO18|OGLUM07G12950
Ory.ba-AGO18|OBART07G13940
Bra.st-AGO18|Brast06G129200.1.p
Bra.hy-AGO18|Brahy.S06G0144700.1.p
Bra.sy-AGO18|Brasy6G154700.1.p
Bra.hy-AGO18|Brahy.D01G0368900.1.p
Bra.di-AGO18|Bradi1g28260.5.p
Thi.in-AGO18|Thint.16G0470400.1.p
Thi.in-AGO18|Thint.08G0142400.1.p
Hor.vu-AGO18|HORVU.MOREX.r2.3HG0197830.1
Sec.ce-AGO18|SECCE3Rv1G0162720.1
Tri.tu.dur-AGO18-3A|TRITD3Av1G037800
Tri.ur-AGO18|TRIUR3_29411
Tri.tu.dic-AGO18-3A|TRIDC3AG014480
Tri.ae-AGO18-3A|TraesCS3A02G116800
Tri.ae-AGO18-3D|TraesCS3D02G118600
Aeg.ta-AGO18|AET3Gv20256400
Thi.in-AGO18|Thint.07G0199100.1.p
Tri.tu.dic-AGO18-3B|TRIDC3BG019000
Tri.tu.dur-AGO18-3B|TRITD3Bv1G046390
Tri.ae-AGO18-3B|TraesCS3B02G135800
BOP clade
Mis.si-AGO18|Misin03G248300.1.p
Zea.ma-AGO18|Zm00001d033206_P001
Sor.bi-AGO18|Sobic.001G153600.1.p
Mis.si-AGO18|Misin02G113300.1.p
Set.vi-AGO18|Sevir.2G321900.2.p
Set.it-AGO18|Seita.2G310800.1.p
Pan.ha-AGO18|Pahal.2G369600.1.p
Pan.vi-AGO18|Pavir.2KG424900.1.p
Pan.vi-AGO18|Pavir.2NG477400.1.p
Era.cu-AGO18|EJB05_50467
Era.cu-AGO18|EJB05_12759
Oro.th-AGO18|Oropetium_20150105_13290A
Ele.co-AGO18|ELECO.r07.7BG0598050.1
Ele.co-AGO18|ELECO.r07.7AG0566640.1
Set.vi-AGO18|Sevir.2G321800.1.p
Pan.ha-AGO18|Pahal.2G369500.1.p
Pan.vi-AGO18|Pavir.2KG424800.2.p
Pan.vi-AGO18|Pavir.2NG477300.4.p
Zea.ma-AGO18|Zm00001d021345_P001
Zea.ma-AGO18|Zm00001d006351_P001
Sor.bi-AGO18|Sobic.002G299200.2.p
Mis.si-AGO18|Misin04G281500.1.p
Mis.si-AGO18|Misin03G248400.1.p
PACMAD clade

Figure 5. AGO proteins in the AGO1/5/10/18 clade.

(A) Pruned tree of AGO1/5/10/18 shows the presence and absence of lineages in each group. The tree shows the plant lineages in which AGOs of these clades diverged. (B) AGO18 tree showing the presence of this AGO in the Cyperaceae and almost every Poaceae species but absent in other monocots. These clades were pruned from a Maximum likelihood tree of 2,991 AGO proteins. Color coding, species codes and names, and bootstrap values are as described in Figure 1. The annotated tree in Newick format is provided in Supplemental Data 3.


A
Duplicated AGO1 copy in PACMAD
Poaceae-AGO1-iv
AGO1 duplication early after the monocot/eudicot split
AGO1 copy absent from in Brachypodieae
Poaceae-AGO1-iii
Poaceae-AGO1-ii
Poaceae-AGO1-i
AGO1 copy absent in maize and sorghum
Tree scale: 1

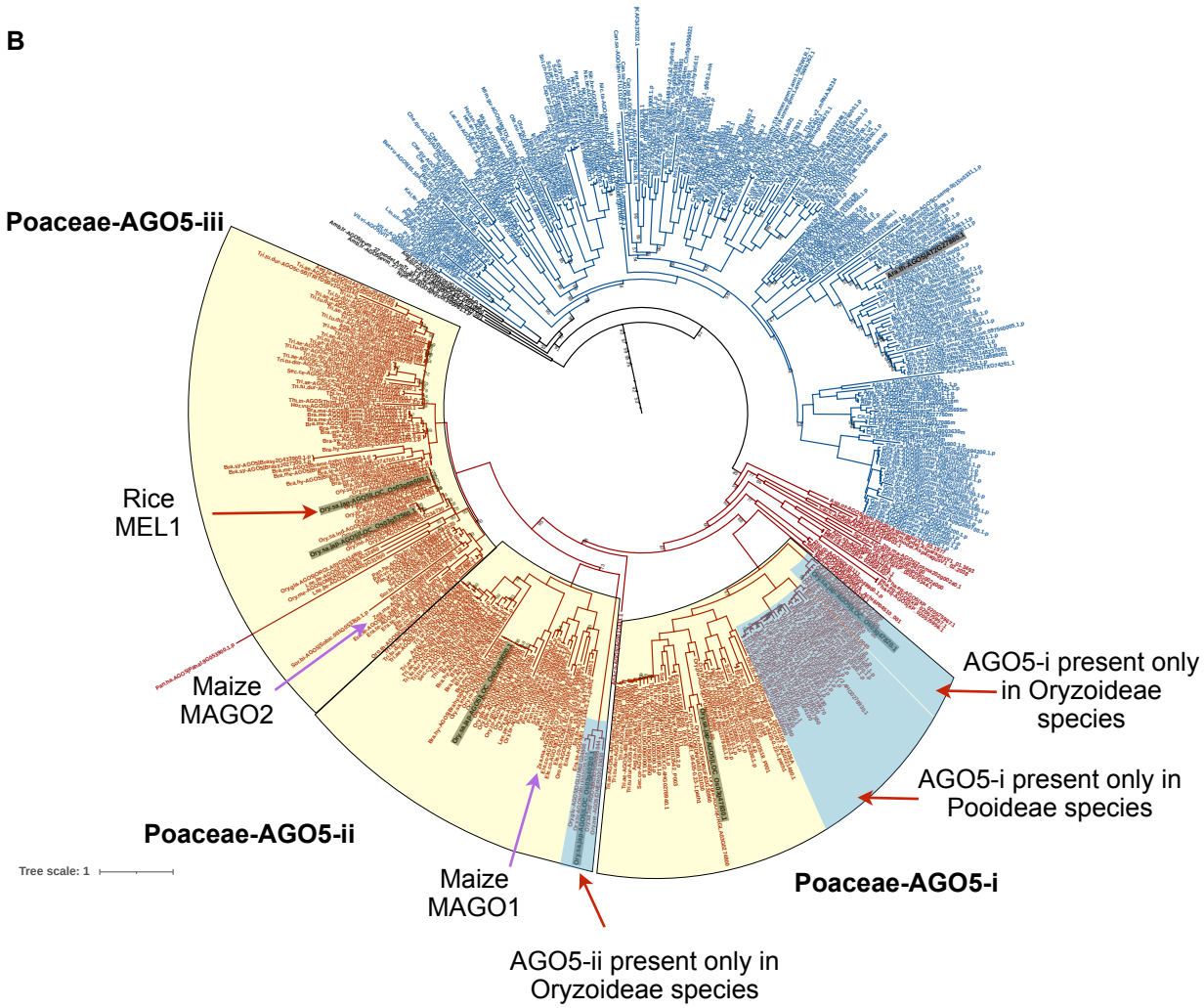

B
Poaceae-AGO5-iii
Rice MEL1
Maize MAGO2
AGO5-i present only in Oryzoideae species
AGO5-i present only in Pooideae species
Poaceae-AGO5-ii
Tree scale: 1
Maize MAGO1
Poaceae-AGO5-i
AGO5-ii present only in Oryzoideae species

Figure 6. AGO proteins in the AGO1 and AGO5 clades.

AGO1 (A) and AGO5 (B) clades pruned from a Maximum likelihood tree of 2,991 AGO proteins. Trees show the plant lineages in which AGOs of these clades diverged. We highlight the absence (shaded blue panel) and presence (shaded green panel) of AGOs specific to monocot subgroups. Color coding, species codes and names, and bootstrap values are as described in Figure 1. The annotated tree in Newick format is provided in Supplemental Data 3.