ANGIOSPERMS ARE UNIQUE AMONGST LAND PLANT LINEAGES IN THE OCCURRENCE OF KEY GENES IN THE RNA DEPENDENT DNA METHYLATION (RdDM) PATHWAY

Lu Ma¹, Andrea Hatlen¹, Laura J. Kelly¹, Hannes Becher¹, Wencai Wang¹, Ales Kovarik², Ilia J. Leitch³ and Andrew R. Leitch¹*

¹School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, UK
²Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno CZ–61265, Czech Republic
³Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, UK

*corresponding author
ABSTRACT

The RNA-directed DNA methylation pathway (RdDM) can be divided into three phases: (i) small interfering RNA biogenesis, (ii) de novo methylation and (iii) chromatin modification. To determine the degree of conservation of this pathway we searched for key genes amongst land plants. We used OrthoMCL and the OrthoMCL Viridiplantae database to analyse proteomes of species in bryophytes, lycophytes, monilophytes, gymnosperms and angiosperms. We also analysed small RNA size categories and, in two gymnosperms, cytosine methylation in ribosomal DNA.

Six proteins were restricted to angiosperms, these being NRPD4/NRPE4, RDM1, DMS3, SHH1, KTF1 and SUVR2, although we failed to find the latter three proteins in *Fritillaria persica*, a species with a giant genome. Small RNAs of 24 nucleotides in length were abundant only in angiosperms. Phylogenetic analyses of Dicer-like (DCL) proteins showed that DCL2 was restricted to seed plants, although it was absent in *Gnetum gnemon* and *Welwitschia mirabilis*.

The data suggest that phases (i) and (ii) of the RdDM pathway, described for model angiosperms, evolved with angiosperms. The absence of some features of RdDM in *F. persica* may be associated with its large genome. Phase (iii) is probably the most conserved part of the pathway across land plants. DCL2, involved in virus defence and interaction with the canonical RdDM pathway to facilitate methylation of CHH, is absent outside seed plants. Its absence in *G. gnemon*, and *W. mirabilis* coupled with distinctive patterns of CHH methylation, suggest a secondary loss of DCL2 following the divergence of Gnetales.

Keywords: chromatin modification, DNA methylation, evolution, RNA dependent DNA methylation, seed plants.
INTRODUCTION

The first land plants appeared in the fossil record around 470-480 million years ago (Mya) (Kenrick and Crane 1997; Wellman et al. 2003) and the species which survive today can be broadly divided into four major groups: (1) the non-vascular plants, which comprise bryophytes (liverworts, mosses and hornworts), (2) the lycophytes, the earliest diverging extant group of vascular plants; (3) the monilophytes which include the horsetails (Equisetum), whisk ferns (e.g. Psilotum), ophioglossoid ferns (e.g. Ophioglossum) and true ferns, and; (4) the seed plants comprising angiosperms (flowering plants) and gymnosperms (naked-seed plants). With the huge increase in genomic data available for species belonging to these different groups, it has become clear that the genome dynamics of each group are distinctive (reviewed in Leitch and Leitch (2012, 2013)). Here we explore the composition of the epigenetic machinery in representatives of these major groups and suggest how the differences encountered might have played a role in shaping their genome dynamics. In particular, we compare and contrast the genes involved in controlling the RNA dependent DNA methylation (RdDM) pathway, with a particular emphasis on angiosperms and gymnosperms, but including representatives of the other land plant groups to determine directionality of change in the evolution of this pathway.

Gymnosperms comprise c. 780 species and are represented by four distinct lineages, the cycads (Cycadales), Ginkgo (Ginkgoales), Gnetales and Coniferales (conifers). Our understanding of their genome structure, and the epigenetic processes that regulate their genomes, is largely restricted to Pinaceae (Leitch and Leitch 2012). Outside this family understanding is minimal, and in most cases missing entirely. Nevertheless, despite this dearth of knowledge, we do know that gymnosperms have reduced frequencies of polyploidy in all but Ephedra (Gnetales) (Leitch and Leitch 2012) and there is some evidence of alternative mechanisms to regulate the evolution of their genome, e.g.
different epigenetic marks associated with heterochromatin (Fuchs et al. 2008), higher levels of transcription of retrotransposons in conifers than angiosperms (Morse et al. 2009; Parchman et al. 2010), and lower levels of unequal recombination to remove the long terminal repeats (LTRs) from LTR retrotransposons (Nystedt et al. 2013). Such differences have been postulated to have fundamentally shaped patterns of genome evolution in seed plants (Leitch and Leitch 2012).

It is widely recognised that the diversity of genome sizes in land plants arises from differences in the accumulation of repetitive elements, including tandem and dispersed repeats, as well as the frequency of polyploidy, or whole genome duplication, in the lineages’ ancestry. This paper focuses on the evolution of mechanisms that control the accumulation of repeats and searches for differences in these mechanisms between representative species of the major land plant lineages.

Regulation of repeats in angiosperms broadly falls into two categories (i) RdDM \textit{de novo} methylation and (ii) maintenance methylation pathways, the latter involving genes which play a role in CG and CHG methylation. This work focuses on the RdDM pathway, leading to the heterochromatinisation of repeats in angiosperms, as summarized in Figure 1, which outlines the canonical pathway. Briefly, the RdDM pathway can be divided into three phases: (i) RNA polymerase IV (Pol IV)-dependent small interfering RNA (siRNA) biogenesis, (ii) RNA polymerase V (Pol V)-mediated \textit{de novo} DNA methylation and (iii) chromatin alteration or modification (review in Matzke and Mosher (2014)). In the first of these, Pol IV activity synthesizes RNA transcripts, which are made double stranded by RNA DEPENDENT RNA POLYMERASE 2 (RDR2) and “diced” or cut into 24 nt siRNAs using DICER-LIKE 3 (DCL3) endonuclease. These siRNAs are then complexed with the ARGONAUTE (AGO) protein AGO4 and directed back to the nucleus. Then the siRNAs, through sequence homology, are targeted back to DNA repeats. In phase (ii) Pol
V is involved in the further transcription of repeats in association with the diced 24 nt siRNA to facilitate RdDM in a little understood process. Finally, in phase (iii), genes involved in histone modification and chromatin folding heterochromatise the DNA sequence. This process then “seeds” methylation, which may spread into surrounding genic regions and become extended and supplemented by the activities of the maintenance methylation pathways which typically involve the recognition and full methylation of hemi-methylated CG and CHG sites by METHYLTRANSFERASE 1 (MET1) and CHROMOMETHYLASE 3 (CMT3) DNA methyltransferases, respectively and of CHH by CMT2 (Matzke and Mosher 2014).

The vast majority of research into the genes involved in the RdDM pathway in plants has been conducted in *Arabidopsis thaliana*. Thus to search for the occurrence of these genes across the different land plant groups, we retrieved the key genes and their paralogues from *A. thaliana* (Figure 2) and used these to search for their occurrence in cluster groups of OrthoMCL from publically available proteome sequence databases of bryophytes, lycophytes, monilophytes, representatives from the gymnosperm lineages, the early diverging angiosperm *Amborella trichopoda* and some model angiosperms (e.g. *Zea mays*). In addition, because gymnosperms have significantly larger genomes than most angiosperms, we hypothesised that this may be due to different activities of RdDM. To test that hypothesis we also searched the transcriptome of *Fritillaria persica*, an angiosperm with a particularly large genome (1C = 41.21 pg) (Kelly et al., 2015).
MATERIALS AND METHODS

Data used to search for orthologues

A flow diagram illustrating our bioinformatic approaches is shown in Figure S1. Twelve proteomes from representative taxa of land plants were selected comprising (i) the angiosperms *Arabidopsis thaliana* L. (Heynh.), *Amborella trichopoda*, Baill. *Fritillaria persica* L., *Zea mays* L.; (ii) the gymnosperms *Ginkgo biloba* L., *Gnetum gnemon* L., *Picea abies* (L.) H.Karst., *Pinus taeda* L., *Welwitschia mirabilis* Hook.f.; (iii) the monilophyte *Pteridium aquilinum* L.Kuhn.; (iv) the lycophyte *Selaginella moellendorffii* Hieron; and (v) the bryophyte *Physcomitrella patens* (Hedw.) Bruch & Schimp.

Proteome data for eight of these species were retrieved from public databases (see Table 1). For the remaining four species we derived proteome data from transcriptomes. For *P. aquilinum* we used the transcriptome data from Li et al. (2014). For *G. gnemon*, we downloaded Illumina raw reads from NCBI SRA archive (ERR364403) (Table 1). For *F. persica* and *W. mirabilis*, we used new transcriptomic data (see below), and reads from the mRNA library of *F. persica*, *G. gnemon* and *W. mirabilis* were *de novo* assembled (Table 1) using Trinity (version r2013-02-25) (Grabherr et al. 2011) with default settings. TransDecoder was then used to identify the protein-coding regions from the *de novo* assembled contigs using default settings and keeping sequences longer than 100 amino acids.

RNA Sequencing

For *F. persica* we obtained transcriptomic data by extracting mRNA from leaves as in Becher et al. (2014). The transcriptome of *F. persica* was sequenced by the Centre of Genomic Research at the University of Liverpool, UK using HiSeq2000 (100 bp paired-
end reads). Total RNA of *W. mirabilis* from fresh leaf fragments was extracted using a mirVana™ miRNA isolation kit (Life Technology, U.S.A.) following the manufacturer’s instructions. Both transcriptome and small RNA sequencing of *W. mirabilis* was conducted by BGI, Shenzhen, China on the HiSeq2000 platform (library fragment size for *W. mirabilis* transcriptome sequencing was 270 bp with 91 bp paired-end reads; library fragment size for *W. mirabilis* small RNA sequencing was 107 bp with 50 bp single-end reads).

**Finding OrthoMCL gene groups of the RdDM pathway in land plants**

Proteomes from the 12 representative land plant taxa were filtered using the pipeline OrthoMCL (v2.0.9) (Li et al. 2003, Fischer et al. 2011) and the number of “good” protein sequences, as defined by OrthoMCL (using default settings), for each species is shown in Table 1. We searched the proteomes against each other using blastp, and then with OrthoMCL. We generated OrthoMCL groups of proteins (clusters) based on similarity, keeping matches with E-values < 1e-5 and ≥ 50% match along the protein length. The MCL algorithm was used to generate the OrthoMCL clusters of proteins with an inflation value of 1.5. To find orthologues of genes in the RdDM pathway we first retrieved the protein sequences listed in Figure 2 (see also Table S1 for the full names of each protein) from *A. thaliana*, and used these to extract orthologous and paralogous proteins from the OrthoMCL clusters of the other 11 species. The protein information retrieved for DMS3, KTF1, DCL and RDM1 is given in Tables S2-S5 respectively, and all protein sequences from each group are given in FASTA format in the Supplementary Data File 1. Custom Python scripts (available on request) were used to extract the protein groups and corresponding protein sequences for each locus in the RdDM pathway based on the gene names used for *A. thaliana* (reference proteins, Table S1).
For six of the RdDM genes, the OrthoMCL groups did not contain sequences from any of the non-angiosperm species analysed. For these genes we also searched for orthologues in the OrthoMCL Viridiplantae database (http://www.orthomcl.org/orthomcl/) by blastp. The OrthoMCL Viridiplantae database includes data from the following eight plant species – angiosperms: *A. thaliana*, *Oryza sativa* and *Ricinus communis*; bryophytes: *P. patens*, and green algae: *Chlamydomonas reinhardtii*, *Micromonas sp. RCC299*, *Ostreococcus tauri* and *Volvox carteri*.

**Generation of Dicer-like (DCL) protein trees**

Using the approach above, DCL putative orthologues from the 12 representative land plant species were extracted from the OrthoMCL output by searching for OrthoMCL groups containing each of the four *Arabidopsis* DCL genes (i.e. DCL1, accession AT1G01040; DCL2, accession AT3G03300; DCL3, accession AT3G43920; DCL4, accession AT5G20320). Protein domains of all sequences were analysed by scanning predicted protein sequences against the Pfam protein database (http://pfam.xfam.org/search). When more than one splice variant was present for a gene, only the longest protein sequence was kept for analysis. When more than one incomplete protein from the same species had the same domains, we kept the longest variant. Protein sequences that passed these selection criteria were aligned using MUSCLE with default parameters (version 3.8.31) (Edgar 2004) and trimmed using trimAl (version 1.2rev59) (Capella-Gutiérrez et al. 2009) with the setting “automated1” to remove regions with an excessive amount of missing data or poorly aligned regions. We used ProtTest (version 3.4) (Darriba et al. 2011) to select the best model (LG+I+G) based on Bayesian Information Criterion (BIC), and RAxML (version 7.4.2) (Stamatakis 2006) to build the phylogenetic trees, performing 1000 bootstrap replicates and using the following options: -p 12345, -f a, -c 4, -x 12345.
Generation of RNA-directed DNA methylation 1 (RDM1) protein tree

Phylogenetic analysis of protein sequences from the RDM1 locus was performed in the same way as described above for DCL. However, only four RDM1 protein sequences were isolated from the OrthoMCL results, all from angiosperms. Consequently, we also searched for putative homologues by performing blastp searches against the NCBI Protein Reference Sequence database (http://www.ncbi.nlm.nih.gov/protein), retaining all protein matches with an E-Value < 1e-5 and ≥ 50% identity.

Small RNA analysis

Most of the small RNA (sRNA) data analysed were downloaded from public databases (Table 1) and comprised reads that had already been trimmed to remove adapter sequences. For W. mirabilis, sRNAs were sequenced here (see above). Custom Python scripts were used to obtain the length of each sRNA sequence within the 18-26 nucleotide (nt) size range.

Southern hybridization

Purified genomic DNAs of G. biloba, G. gnemon and, as a control, the angiosperm Nicotiana tabacum L. (~2 μg/sample) were digested with the restriction enzymes MspI, HpaII, BstNI or ScrFI and separated by gel electrophoresis on a 0.9% (w/v) agarose gel. The gels were then alkali-blotted onto Hybond-XL membranes (GE Healthcare, Little Chalfont, UK) and hybridized with a 32P-labelled DNA probe (DekaLabel kit, MBI, Fermentas, Vilnius, Lithuania) for the 18S ribosomal RNA (rRNA) gene according to protocols described in Kovarik et al. (2005). After washing (2 x 5 min with 2x SSC, 0.1%
SDS at room temperature followed by 2 x 15 min with 0.6 x SSC, 0.1% SDS, 65°C), the hybridization bands were visualised with a PhosphorImager (Typhoon 9410, GE Healthcare, PA, USA) and the data quantified by ImageQuant software (GE Healthcare, PA, USA). The 18S probe was a 300-bp fragment (Figure 6a) obtained by amplification of the 18S rRNA gene of the gymnosperm *Cycas revoluta* Thunb. using primers described further below.

**Bisulphite sequencing**

Modification of DNA with bisulphite was carried out with an EpiTect kit (Qiagen, Germany) using 1.3 μg of genomic DNA from leaves. The primers used amplified the coding strand of the 18S rRNA gene subregion shown in Figure 6a and did not discriminate between methylated and non-methylated templates. The primer sequences were as follows: 18SBIS forward: 5’-TATGAGTYTGTAATTGGAATG-3’; 18SBIS reverse: 5’-TTTAARCACTCTAATTTCTTCAAA-3’. The PCR reaction (total volume 25 μl) used 1.0 μl of bisulphite-converted DNA as the template, 4 nmol of each dNTP, 8 pmol of each primer and 0.8 U of Kapa *Taq* DNA polymerase (Kapabiosystems, USA). Cycling conditions were as follows: initial denaturation (94°C/3 min); 35 cycles of (94°C/20 s; 55°C/20 s; 72°C/20 s) and a final extension (72°C/10 min). Resulting ~300 bp products were separated by gel electrophoresis, purified and cloned into a TA vector (pDrive, Qiagen). In total, 22 and 18 clones were sequenced from *G. gnemon* and *G. biloba*, respectively. After trimming of primers the 241 bp-long sequences were aligned and statistically evaluated using CyMATE software (Hetzl et al. 2007).

**RESULTS**
OrthoMCL clustering

The proteomes of 12 species were compiled to include representative taxa from all four major land plant lineages. Together these 12 taxa generated between 18,255 (W. mirabilis) and 88,760 (Z. mays) protein sequences, summing to a total of 543,399 proteins that were clustered into 55,357 OrthoMCL groups (containing both paralogues and orthologues) using OrthoMCL (Table 1).

We found OrthoMCL groups for all 31 genes/gene families listed in Figure 2 which represent genes belonging to the three phases of the canonical RdDM pathway, namely: (i) Pol IV-dependent siRNA biogenesis, (ii) Pol V-mediated de novo DNA methylation and (iii) chromatin alterations (Figure 1) together with additional factors also involved in cytosine methylation. OrthoMCL groups of nine proteins or families involved in the RdDM pathway, namely the NRPD2/NRPE2, NRPE9B, NRPB1, RDR, DCL, HEN, AGO, HDA, and UBP contained sequences from all 12 of the species analysed, indicating high levels of conservation for these loci across land plants (highlighted in green Table S1). MET1, which codes for METHYLTRANSFERASE 1, and DDM1 (DECREASED DNA METHYLATION 1), which is a chromatin remodeller protein, were also found in all analysed species.

Putative homologues of DMS3 (DEFECTIVE IN MERISTEM SILENCING 3) were found in all plants except P. taeda and P. patens (Table S2). However, closer analysis revealed that the protein was either unusually long, indicative of SMC proteins involved in chromatin remodelling (Matzke et al. 2014), or so short that it was not possible to distinguish DMS3 from SMC homologues. Consequently proteins >700 amino acids and <150 amino acids were removed. Proteins with a histidine kinase-like ATPase motif, present in the SMC-related protein AtGMI1 (Böhmdorfer et al. 2011) but not in DMS3, were also removed. This left only putative DMS3 OrthoMCL groups in seed plants. These
proteins were aligned using T-Coffee (Notredame et al. 2000) and the alignment quality was assessed using Transitive Consistence Score (TCS, Chang et al. 2015). Four proteins had poor alignment (TCS ≤ 16, see Table S2) and these, together with one isoform of the protein from *Zea mays*, were removed from the analysis, leaving seven sequences, all from seed plants. The SMC-related protein from *A. thaliana* (GMI1_AT5G24280) was added to the alignment. Phylogenetic analysis of these eight sequences revealed two groups, one containing angiosperms, the other gymnosperms, each being separated by GMI1 from *A. thaliana* (Figure S2). Thus, sequences from gymnosperms cannot be distinguished from SMC-related proteins, and only in angiosperms can we confidently identify DMS3-like sequences, consistent with Matzke et al. (2015).

We searched the data for OrthoMCL groups that were found only in angiosperms and so missing in all other land plant groups and found, in addition to DMS3, a further five proteins in this category: NRDP4/NRPE4, SHH1, RDM1, SUVR2 and KTF1 (all shown in bold in Figure 2 and highlighted in blue in Table S1). All six proteins belong to phase (i) and/or (ii) of the RdDM pathway. From the 12 proteomes included in the OrthoMCL analysis, only the eudicot *A. thaliana* yielded sequences for SUVR2 and SHH1 (but see below).

We extended our proteome analysis to include the OrthoMCL Viridiplantae database, which contains data from six additional plant species not analysed above (*Oryza sativa, Ricinus communis, Chlamydomonas reinhardtii, Macromonas sp., Ostreococcus tauri* and *Volvox carteri*). We focused our search on identifying homologues of the six proteins found only in angiosperms (see above). Using this extended approach SUVR2 and SHH1 were found in all three angiosperm species listed in the OrthoMCL Viridiplantae database, including the monocot *O. sativa*, showing that these gene families are not restricted to eudicots. In *A. thaliana* we identified three putative homologues of the SHH1-family and
five of SUVR-family, whilst in both *O. sativa* and *R. communis* we identified one putative homologue in each.

Beyond the angiosperms, no sequences with homology to NRPD4/NRPE4, SHH1, RDM1, and SUVR2 were found in gymnosperms, monilophytes or lycophytes, but putative KTF1 (KOW DOMAIN-CONTAINING TRANSCRIPTION FACTOR 1; a synonym of SPT5L) homologues were found in the bryophyte *P. patens*, and the green algae *Micromonas* sp. RCC299 and *Volvox carteri* in the OrthoMCL Viridiplantae database. Because OrthoMCL relies on low thresholds of blast similarity (E-values < 1e-5 and ≥ 50% match along the protein length), we further characterised these proteins, by searching for NGN and KOW domains, together with the WG/GW motifs characteristic of KTF1 (He et al. 2009b, Matzke et al. 2015) (Figure S3). We failed to find NGN and KOW domains outside the angiosperms (Table S3). We also noticed that while the putative KTF1 sequences in the angiosperm *F. perisca* contained both NGN and KOW domains (Matzke et al. 2015), they lacked GW/WG motifs, perhaps because the protein is a partial assembly (Figure S3, Table S3).

In summary, the combined results from our analysis indicated that NRPD4/NRPE4, SHH1, RDM1, KTF1, DMS3 and SUVR2 are restricted to angiosperms.

**Phylogenetic relationships between members of the Dicer-like (DCL) family proteins**

In *A. thaliana* it is known that there are four paralogues in the DCL family, these are DCL1 which generates 21 nt microRNAs (miRNAs), DCL2 generating 22 nt siRNAs from viral sequences, DCL3 involved in RdDM and generating 24 nt siRNAs (Figure 1) and DCL4, generating 21 nt siRNAs and trans-acting siRNAs (tasiRNA). In *A. thaliana*, expression levels of each DCL family member are similar and at medium levels in most tissues.
(Zimmermann et al. 2004), so we might expect to detect the presence of orthologues in other species, if they are present.

From the 12 species that are the focus of this study, a total of 84 proteins formed a “DCL family group”. They included the four DCL family members in *A. thaliana* (DCL1 – DCL4), which, when complete, should each exceed 1,300 amino acids. Protein domains of all sequences were analysed by scanning against the Pfam protein database ([http://pfam.xfam.org/search](http://pfam.xfam.org/search)). When more than one splice variant, or size variant was present, only the longest protein sequence was kept for analysis. All sequence variants were kept, leaving 56 protein sequences (Table S4). Phylogenetic relationships between the DCL family members are shown in Figure 3. The sequences group into four strongly supported clades, and in each clade there is an *A. thaliana* DCL member, as expected. This enabled us to label the clades DCL1 to DCL4. Recently, it has been noted that there are two distinct clades of DCL3-like sequences in monocots, called DCL3a and DCL3b, the later renamed DCL5 (Margis et al. 2006, Song et al. 2012, Fei et al. 2013), and represented by two DCL3 clades each containing *Z. mays* and *O. sativa* sequences (see Figure 3). Table 2 summarises the number of sequences (paralogues) for each of the four DCL family members across the 12 species analysed. All species had proteins related to DCL1. Of particular note was DCL2, which was absent outside the seed plants (i.e. angiosperms and gymnosperms) and, perhaps significantly, also absent in the two gymnosperms analysed belonging to Gnetales (*W. mirabilis* and *G. gnemon*). There were also isolated absences of DCL3 (in *W. mirabilis* and *P. aquilinum*) and DCL4 (in *G. gnemon* and *S. moellendorffii*). It may also be significant that we found only two DCL4 domains in *F. persica* (Helicase C and Dicer-dimer domains, out of the nine DCL domains considered, Table S4).

**RNA-directed DNA methylation 1 (RDM1) family**
OrthoMCL clustering revealed one RDM1 orthologue each for *A. thaliana*, *F. persica*, *Z. mays* and *A. trichopoda*. To better understand the evolution of RDM1, we BLAST-searched the NCBI Protein Reference Sequence database to look for further sequences with similarity to RDM1 and found 68, all from 35 angiosperm species (comprising one early-diverging, 3 monocot, and 12 eudicot families; Table S5). These were aligned and used to build a phylogenetic tree of sequence relationships (Figure 4). Within the RDM1 phylogenetic tree, three family-specific clades were recovered; one comprising all the sequences from Brassicaceae species, another containing all the sequences from Fabaceae species and a third made up of sequences from species belong to Solanaceae. These three eudicot clades were very strongly supported (bootstrap support >95%). For five genera from four further eudicot families with two or more sequences (i.e. *Citrus* (Rutaceae), *Cucumis* (Cucurbitaceae), *Theobroma* (Malvaceae), and *Fragaria* and *Pyrus* (both Rosaceae)), the sequences clustered by genus with strong support. A further clade was identified which contained all RDM1 sequences from monocot species, but it lacked strong support.

**Length distribution of small RNA (sRNA)**

Because the total number of available sequences differed in the eight sRNA datasets examined (from angiosperms: *A. trichopoda*, *Z. mays*, *A. thaliana*; from gymnosperms: *G. biloba*, *P. abies*, *W. mirabilis*; from lycophytes: *S. moellendorkffii*; and from bryophytes *P. patens*) (see Table 1), we plotted the percentage of sRNA sequences belonging to each size category (Figure 5). The most abundant category was 24 nt for all angiosperms (Figure 5a). In contrast, for all other land plants analysed the 21 nt sRNA size category was most abundant (Figure 5b).
Cytosine methylation in the gymnosperms *Ginkgo biloba* and *Gnetum gnemon*

Since DCL2 was missing in *G. gnemon* and *W. mirabilis*, and it is thought to interact with RdDM in the non-canonical methylation of cytosine (involving RDR6) (Nuthikattu et al. 2013), we conducted bisulphite sequencing of the 18S rDNA in *G. gnemon* and *Ginkgo biloba* to compare levels of CHH methylation. Figure 6b shows that in both species CG and CHG methylation levels were high, but CHH methylation was very low in *G. gnemon*.

To further study methylation patterns, we used Southern hybridization and an 18S rDNA probe against restricted genomic DNA (using methylation-sensitive and insensitive isoschizomers) from the gymnosperms *G. gnemon*, *G. biloba* and the angiosperm *Nicotiana tabacum* (chosen as a control because the methylation status of its rDNA has been extensively studied; Lim et al. 2000). We revealed more extensive digestion of *G. gnemon* DNA with *Msp*I (sensitive to CHG methylation) compared with the other species (Figure S4, red circles). This confirmed a relative undermethylation of *G. gnemon* rDNA. In contrast, the fraction of rDNA resistant to digestion with methylation-sensitive enzymes (Figure S4, red bars) was relatively high in *G. biloba* and *N. tabacum*, indicating dense methylation of their units.

**DISCUSSION**

Differences in RdDM pathway genes across land plants

Six proteins were identified in our analyses that were absent or missing outside angiosperms (Figure 7), these are NRPD4/NRPE4, SHH1, RDM1, DMS3, KTF1 and SUVR2. All are involved in phases (i) and (ii) of the canonical RdDM pathway (Figures 1, 2 and Table S1). Collectively, the data suggest that phases (i) and (ii) of the RdDM pathway have diverged between the different land plant groups whereas phase (iii), which
Ma et al. (2015) RdDM pathway in land plants

is involved in chromatin remodelling, is the most highly conserved part of the pathway. The other proteins of phases (i) and (ii) that we analysed were found across land plants, perhaps with variant functions outside the seed plants (as in DCL, see below).

NRPD4/NRPE4 is known to function as part of the RNA polymerase IV (Pol IV) and V (Pol V) complexes. It is encoded by the same gene and is distinct from the NRPB4 subunit of RNA polymerase II (Pol II) in *Arabidopsis thaliana* (He et al. 2009a; Ream et al. 2009). NRPD4/NRPE4 forms sub-complexes with NRPD7 and NRPE7 in Pol IV and Pol V respectively (Ream et al. 2009). Pol IV and Pol V are central to the RdDM pathway and probably to its evolution (Matzke et al. 2015). Previously it was suggested that NRPD4 evolved after the divergence of *P. patens* and before angiosperms (Tucker et al. 2010). Our data extend these findings by showing that NRPD4/E4 diverged with the angiosperms.

Pol IV is thought to be recruited to a subset of target loci for siRNA production by the protein SHH1 (SAWADEE HOMEODOMAIN HOMOLOG 1) which recognizes and binds to H3 histones when they are unmethylated at lysine 4 (=H3K4) and methylated at lysine 9 (=H3K9), i.e. markers of heterochromatin production (Law et al. 2013; Matzke et al. 2015; Zhang et al. 2013a). Our failure to detect SHH1 outside angiosperms is consistent with the lack of the NRPD4/NRPE4 subunits of Pol IV.

RDM1 (RNA-DIRECTED DNA METHYLATION 1) is reported to be needed for Pol V function (Matzke et al. 2015) and is currently understood to interact with the Pol V pathway in phase (ii) of the RdDM pathway in two ways: (a) acting as a homodimer protein bridging between AGO4 and DRM2 in the de novo methylation step (Gao et al. 2010; Sasaki et al. 2014), and (b) acting as a monomer in the DDR complex (together with DRD1 and DMS3) that facilitates Pol V transcription (Law et al. 2010). Certainly, Arabidopsis *rdm1* mutants show a nearly complete loss of DNA methylation via the RdDM pathway (Stroud et al. 2013;
Gao et al. 2010; Sasaki et al. 2014). Previously Matzke et al. (2015) noted that RDM1 was restricted to angiosperms, and we confirm this in our taxonomically more diverse survey, which includes representatives from all major land plant groups. The phylogenetic tree inferred from RDM1 sequences from 35 angiosperm species illustrates that sequences from three eudicot families cluster into discrete, highly supported, clades (Figure 4).

Overall, it appears that the specialised components of both Pol IV and Pol V pathways may only be present in angiosperms (Figure 7).

The protein SUVR2 was also shown to be restricted to angiosperms (Figure 2 and Table S1). It is a putative histone methyltransferase that is not directly required for the generation of siRNAs by the RdDM pathway, but was recently shown to be required for DRM2 establishment and for maintaining methylation downstream of siRNA biogenesis (Stroud et al. 2013).

The final angiosperm specific protein is KTF1, a transcription factor that plays a role in phase (ii) of the RdDM pathway by coordinating transcriptional elongation with chromatin modifications and pre-mRNA processing via interactions with AGO4 (He et al. 2009b). It was previously reported to be restricted to angiosperms (Matzke et al. 2015), consistent with findings from the more extensive survey here.

**DICER-like (DCL) proteins**

DCL proteins are multi-domain endoribonucleases, which ‘dice’ or cut pre-matured long double stranded RNAs into sRNAs (Bernstein et al. 2001; Liu et al. 2009; Matzke et al. 2015). The number of DCL family members varies among different organisms and patterns of evolution across eukaryotes, including an alga and three angiosperms, have
been discussed previously (Margis et al. 2006). In *A. thaliana* there are four DCL gene paralogues (DCL1 – DCL4) (Schauer et al. 2002), but in other eukaryotic groups the numbers can vary from one to more than four types (Bernstein et al. 2001; Liu et al. 2009). Our analysis showed that only DCL1 was found in all the land plant lineages examined, which suggests it is the most highly conserved. In *A. thaliana*, this protein has a role in generating 21 nt miRNAs involved in post-transcriptional regulation of their target genes.

The isolated absences of other DCL family members in our analysis (e.g. DCL3 in *W. mirabilis* and *P. aquilinum* and DCL4 in *G. gnemon* and *S. moellendorffii*) may have arisen because the gene transcripts were not sequenced or detected by us. We have therefore put more weight on our findings where there is strong phylogenetic signal in the patterns of gene losses and gains (Figure 3).

In our analysis, DCL2 was only detected in species belonging to the seed plants (Figure 3), although it was not found in the two species of the gymnosperm order Gnetales examined (i.e. *W. mirabilis* and *G. gnemon*; Figure 3). It is therefore possible that DCL2 sequences have been secondarily lost with the divergence of these species in Gnetales (Figure 7). DCL2 is thought to be involved in RNA-mediated virus resistance and is associated with the production of 22 nt sRNAs. There may also be interactions between the post-transcriptional gene silencing (PTGS) pathway that targets RNA polymerase II-transcribed genes, including newly transposed retroelements, and the non-canonical methylation of cytosines in the RdDM pathway (Nuthikattu et al. 2013). The latter involves the activities of DCL2 and DCL4 to generate 21 and 22 nt sRNAs. Bisulphite sequencing of *G. gnemon* revealed unusually low levels of CHH methylation in 18S rDNA sequences compared with *G. biloba*, which does have DCL2 (Figure 6b, Figure S4). Such a result is consistent with an absence of interaction of DCL2 with RdDM in *G. gnemon*. If so, the absence of
DCL2 outside the seed plants could have similar consequences on the degree of methylation at non-canonical cytosines.

DCL3, which generates 24 nt sRNAs and is directly involved in the canonical RdDM pathway (Figure 1), was found in all plant groups except in the monilophyte studied (Figure 3). Although previous studies failed to detect 24 nt sRNAs in conifers (Dolgosheina et al. 2008), recently they were reported to be present in some tissues of Cunninghamia lanceolata (Wan et al. 2012), P. abies (Nystedt et al. 2013) and Larix leptolepis (Zhang et al. 2013b), consistent with the results presented here. Indeed our survey of the sRNAs generated across land plants shows that all species have a fraction of sRNAs that are 24 nt long although it is only in the angiosperms that these comprise the major fraction of sRNAs (Figure 5). The observation that 24 nt sRNAs were present in the monilophyte P. aquilinum (Figure 5) may indicate that we have simply failed to find DCL3 in the transcriptome data currently available, rather than the gene being absent from their genomes.

DCL4 is thought to be involved in trans-acting RNA metabolism and post-transcriptional gene regulation, generating 21 nt sRNAs. We found DCL4 in all land plant lineages except the lycophyte studied (Table 2 and Figure 3).

In consideration of missing genes in the pathway it must be noted that there is redundancy in function between these DCL families, which results in limited phenotypes in knock-out experiments (Andika et al. 2015). This means that the losses of particular DCL families may be functionally compensated for by the activity of another DCL family member.

Influence of different epigenetic machinery on genome structures
The primary role of RdDM is considered to be the epigenetic silencing of repeats, predominantly retroelements across the genome. This silencing process leads to chromatin remodelling or heterochromatinisation, which typically renders the repeats transcriptionally silent (Matzke et al. 2015; Matzke and Mosher 2014). For example, amongst angiosperms it is known that modifications to, or breakdown of, the RdDM pathway can lead to repeat amplification, as shown for example by the inactivity of an orthologue of RDR2 in *Z. mays* resulting in enhanced transposon activity (Woodhouse *et al.*, 2006).

The differences in the epigenetic machinery amongst representatives of the major land plant groups we show here might potentially influence the evolutionary dynamics of their genomes. Angiosperms are thought to have dynamic genome structures compared with gymnosperms, with a higher level of turnover of retroelements (Leitch and Leitch 2012), at least in those species with a small genome (cf. Kelly *et al.* 2015). Angiosperms are also remarkable amongst comparably-sized eukaryotic groups in terms of their genome size diversity. Not only do they have the largest range for any comparable group – varying c. 2,400-fold (0.063 – 152.23 pg), but the distribution of genome sizes is skewed towards small genomes, with the modal and median values being just 1C = 0.6 pg and 2.5 pg respectively (Leitch and Leitch 2013).

To determine if angiosperms with large genomes have anything unusual in their RdDM pathway we analysed *F. persica*, which has an extraordinary large genome size for any eukaryote (1C = 41.21 pg, Kelly *et al.*, 2015), nearly 300 times that of *A. thaliana*.

Previously, in a study of a related species (*F. imperialis*; 1C = 43 pg), we identified a pararetrovirus-like repeat sequence (FriEPRV) which was estimated to be present in c. 21,000 copies, accounting for 0.4% of its genome (Becher *et al.* 2014). We showed high levels of cytosine methylation and an abundance of 24 nt sRNA reads that mapped
exclusively to the repeat, a result which did not suggest anything unusual in the RdDM pathway. Nevertheless, here we failed to detect NRPD1, SUVR2, and SHH1 (Table S1) in *F. persica*. Potentially KTF1 is also missing since the OrthoMCL group protein isolated lacks GW/WG motifs, which function to interact with AGO4 and siRNAs (He et al. 2009b). However for this protein we cannot rule out incomplete assembly (Figure S3). Similarly we only found two domains for DCL4 (Table S4), although this too might point towards an incomplete assembly. Nevertheless, collectively, it remains possible that there is divergence in particular components of the RdDM pathway in *Fritillaria*, which perhaps impacts on the amplification and elimination of different types of repeat in the genome. If so, this may contribute to the observation that *Fritillaria* genomes comprise a high diversity of highly heterogeneous repeats, each representing a rather small proportion of the genome (Kelly et al. 2015). Such a pattern of repeats may also be present in other species with large genomes (Metcalfe and Casane 2013). This pattern differs from that generally found in species with small genomes, where amplification of one or a few repeat families can result in the contrasting genome sizes observed (Grover and Wendel 2010; Bennetzen and Wang 2014).

In contrast to angiosperms, gymnosperms have relatively limited genome size variation (just 16-fold overall, 2.25 – 36.00 pg) despite having the highest proportion of species with recorded DNA C-values (c. 25% of species) (Leitch and Leitch 2012, 2013). In addition, the mode and median genome size values are significantly higher compared with angiosperms (gymnosperm mode 1C = 10.0 pg, median 1C = 7.9 pg and mean 1C = 18.6 pg). Such differences, coupled with the heterogeneous repeat profiles of the Coniferales species examined (Kovach et al. 2010; Nystedt et al. 2013), could also be related to differences observed in the epigenetic machineries. Potentially in angiosperms RdDM pathways evolved as another, or alternative, layer of transposon proliferation.
control not found in other land plant groups. In angiosperms, it is thought that activated transposons (transcribing RNA) are re-silenced through RdDM. However, we are unaware of evidence for an active transposon in gymnosperms, despite their large genomes, whilst there are many examples in angiosperms (Lisch 2013). Possibly gymnosperms and other land plants have other/alternative mechanisms to silence transposons, such as an elevated frequency of C to T mutation of non-coding, highly methylated repeats.

Available cytological data in monilophytes, lycophytes and bryophytes point to further differences with seed plants in patterns of genome organization (Leitch and Leitch 2013). Sadly, however, the lack of extensive genomic data for these land plant groups precludes generalizations about their genome dynamics and the role that epigenetics may play. It is clear that more molecular studies are needed to probe the role of RdDM in contributing to the contrasting genomic profiles observed across land plants.

ACKNOWLEDGEMENTS

We are grateful for the FP7 Marie Curie IEF (Lu Ma), the FP7 Marie Curie ITN INTERCROSSING (Andrea Hatlen), the Czech Science Foundation (501/12/G090, Ales Kovarik), China Scholarship Council (Wencai Wang), and NERC (NE/ G01724/1, Laura Kelly, Ilia Leitch, Andrew Leitch). The Illumina sequencing of Fritillaria was funded by NERC (NE/G01724/1) and generated by the Centre of Genomic Research in the University of Liverpool, UK. This research utilised Queen Mary’s MidPlus computational facilities, supported by QMUL Research-IT and funded by EPSRC grant EP/K000128/1. We thank an anonymous referee for a rigorous, insightful and helpful review.
REFERENCES


Kovach A et al. 2010. The Pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics. 11:420.


FIGURE LEGENDS

Figure 1. The RdDM pathway, taken from Matzke and Mosher (2014). The genes involved are shown, and details are given, in the source reference. The pathway is divided into three key phases, (i) Pol IV-dependent siRNA biogenesis, (ii) Pol V-mediated de novo DNA methylation, and (iii) chromatin alterations. An overview of their activity is given in the introduction and the full names of genes given in Table S1.

Figure 2. Key genes of the RdDM pathway taken from Matzke and Mosher (2014) and Matzke et al. (2015) that have been analysed. Genes given in bold were not detected by us outside angiosperms. The genes are grouped into three categories in line with the three phases of chromatin remodelling shown in Figure 1.

Figure 3. Phylogenetic relationships between Dicer-like (DCL) sequences showing four distinct DCL clades (DCL1-4). The DCL3b (or DCL5, Song et al. 2012, Fei et al. 2013) clade is labelled with a red asterisk. The (+/-) symbols indicate the land plant group in which each DCL parologue was found. Abbreviations are Physcomitrella patens (PPAT), Selaginella moellendorffii (SMOE), Pteridium aquilinum (PAQU), Pinus taeda (PTAE), Picea abies (PABI), Welwitschia mirabilis (WMIR), Gnetum gnemon (GMON), Ginkgo biloba (GBIL), Amborella trichopoda (ATRI), Fritillaria persica (FPER), Zea mays (ZMAY), Oryza sativa (OSAT) and Arabidopsis thaliana (ATHA).

Figure 4. Unrooted phylogenetic tree depicting relationships between RNA DIRECTED DNA METHYLATION 1 (RDM1)-like protein sequences from angiosperms. All protein sequences used to build the tree were extracted from the NCBI Protein Reference Sequence database by blastp (see Table S5 for Genbank accession numbers used).

Abbreviations are: Amborella trichopoda (A. trich), Arabidopsis lyrata subsp. lyrata (A. lyrat), Arabidopsis thaliana (A. thali), Brachypodium distachyon (B. dista), Brassica rapa (B. rapa), Camelina sativa (C. sativ), Capsella rubella (C. rubel), Cicer arietinum (C.
Lu Ma et al. (2015) RdDM pathway in land plants

ariet), Citrus clementine (C. cleme), Citrus sinensis (C. sinen), Cucumis melo (C. melo),
Cucumis sativus (C. sativu), Eucalyptus grandis (E. grand), Eutrema salsugineum (E.
salsu), Fragaria vesca subsp. vesca (F. vesca), Glycine max (G. max), Medicago
truncatula (M. trunc), Morus notabilis (M. notab), Musa acuminata subsp. malaccensis
(M. malac), Nelumbo nucifera (N. nucif), Nicotiana sylvestris (N. sylve), Nicotiana
tomentosiformis (N. tomen), Phaseolus vulgaris (P. vulga), Phoenix dactylifera (P. dacty),
Populus trichocarpa (P. trich), Prunus mume (P. mume), Prunus persica (P. persi), Pyrus
x bretschneideri (P. brets), Ricinus communis (R. commu), Setaria italica (S. itali),
Solanum lycopersicum (S. lycop), Solanum tuberosum (S. tuber), Theobroma cacao (T.
cacao), Vitis vinifera (V. vinif), Zea mays (Z. mays). Where there were multiple sequences
from a single species, a number follows the taxon abbreviation. Numbers on branches
show bootstrap support values for key nodes discussed in the text; due to reasons of space,
the support values for other nodes have been omitted.

Figure 5. Length distribution of sRNA sequences from (a) three angiosperm species and (b)
five other land plant species listed in Table 1. The percentage of total reads for each size class
is plotted.

Figure 6. Bisulphite sequencing of part of the 18S rDNA in Ginkgo biloba and Gnetum gnemon
was used to determine the level of C methylation (a) Diagrammatic scheme of the G. gnemon
18S rDNA unit (Genbank accession number U42416.1) showing the loop regions (V1-V7,
brown arrows) and the region selected for bisulphite sequencing (blue line). (b) Results of
methylation analysis. Note the relatively low level of non-CG methylation in G. gnemon where
only 4/451 CHH sites (0.9%) were methylated.

Figure 7. Synthesis of data showing likely origin of gene families associated with the
RdDM pathway in the evolution of land plants. The summary tree topology was based on
Mathews (2009).
Table 1. Transcriptomes, proteomes and small RNA (sRNA) data used in this study

<table>
<thead>
<tr>
<th>Species</th>
<th>Abbreviation</th>
<th>Source of proteomes</th>
<th>No. of proteins</th>
<th>Source of sRNA</th>
<th>Tissue for sRNA</th>
<th>No. of sRNAs (18-26 nt)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ANGIOSPERMS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amborella trichopoda</td>
<td>ATRI</td>
<td>Phytozone v10 (<a href="http://phytozome.jgi.doe.gov/">http://phytozome.jgi.doe.gov/</a>)</td>
<td>26,846</td>
<td><a href="http://smallrna.udel.edu/data.php">http://smallrna.udel.edu/data.php</a></td>
<td>Leaves</td>
<td>4,003,853</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>ATHA</td>
<td>Phytozone v10 (<a href="http://phytozome.jgi.doe.gov/">http://phytozome.jgi.doe.gov/</a>)</td>
<td>35,386</td>
<td>GEO (GSM154370)</td>
<td>Leaves</td>
<td>15,831</td>
</tr>
<tr>
<td>Fritillaria persica</td>
<td>FPER</td>
<td>Trinity_assembled (see link*)</td>
<td>62,452</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>GYMNOSPERMS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gnetum gnemon</td>
<td>GMON</td>
<td>Trinity-assembled (see link*) ERR364403</td>
<td>26,782</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Welwitschia mirabilis</td>
<td>WMIR</td>
<td>Trinity-assembled (see link*) See link*</td>
<td>18,255</td>
<td>See link*</td>
<td>Leaves</td>
<td>56,649,017</td>
</tr>
<tr>
<td>Pinus taeda</td>
<td>PTAE</td>
<td><a href="http://loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/">http://loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/</a></td>
<td>64,809</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>MONILIOPHYTES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plant Species</td>
<td>Code</td>
<td>Assembly/Database</td>
<td>Length</td>
<td>GEO Accession</td>
<td>Type</td>
<td>Length</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------</td>
<td>------------------------------------</td>
<td>--------</td>
<td>---------------</td>
<td>-----------------------</td>
<td>---------</td>
</tr>
<tr>
<td><em>Pteridium aquilinum</em></td>
<td>PAQU</td>
<td>NCBI Transcriptome Shotgun Assembly (GASP001)</td>
<td>23,332</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Selaginella moellendorffii</em></td>
<td>SMOE</td>
<td>Phytozome v10 (<a href="http://phytozome.jgi.doe.gov/">http://phytozome.jgi.doe.gov/</a>)</td>
<td>22,285</td>
<td>GEO (GSM176654)</td>
<td>Above-ground tissues</td>
<td>1,302,40</td>
</tr>
<tr>
<td><em>Physcomitrella patens</em></td>
<td>PPAT</td>
<td>Phytozome v10 (<a href="http://phytozome.jgi.doe.gov/">http://phytozome.jgi.doe.gov/</a>)</td>
<td>42,392</td>
<td>GEO (GSM115095)</td>
<td>Proto-nemata</td>
<td>97,999</td>
</tr>
</tbody>
</table>

* [https://goo.gl/PrNKfB](https://goo.gl/PrNKfB)
Table 2. Numbers of paralogues in each of the Dicer-like (DCL) family members

<table>
<thead>
<tr>
<th>Species</th>
<th>Total number</th>
<th>DCL1</th>
<th>DCL2</th>
<th>DCL3</th>
<th>DCL4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ANGIOSPERMS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Zea mays</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Fritillaria persica</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Amborella trichopoda</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>GYMNOSPERMS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ginkgo biloba</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Picea abies</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pinus taeda</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2*</td>
</tr>
<tr>
<td>Gnetum gnemon</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Welwitschia mirabilis</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>MONILOPHYTES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pteridium aquinum</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>LYCOPHYTES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selaginella moellendorffii</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>BRYOPHYTES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physcomitrella patens</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

*The assembled proteins are incomplete, and based on their sequences and domains present (see Table S4) they may represent a single protein.
(i) Pol IV-dependent siRNA biogenesis
(ii) Pol V-mediated de novo DNA methylation
(iii) Chromatin alterations

RNA methylation
DNA methylation
Histone methylation

Dicing
HEN1
AGO4

Pol IV
RDR2
Pol IV transcript
CLSY1
SHH1

Methylation

Pol V
DRM2
Pol V transcript

RDM1
RDM1
RDM1
RDM1

CTD
KTF1

DRD1
DMS3
MORC6

SUHV12/9

IDN2-IDP complex
SWI/SNF complex
HDA6
UBP26

Direction of transcription
Direction of transcription

SMBE
Fig. 2

Phase (i) Pol IV-dependent siRNA biogenesis

- NRPD1
- NRPD2/NRPE2
- NRPD4/NRPE4
- RDR2
- CLSY1
- SHH1
- DCL3
- HEN1
- AGO4
- AGO6
- AGO9

Phase (ii) Pol V-dependent de novo methylation

- NRPE1
- NRPD2/NRPE2
- NRPD4/NRPE4
- NRPE5
- NRPE9B
- AGO4
- DMS4
- KTF1
- RDM1
- DRM2
- DRD1
- MORC1
- DMS3
- MORC6
- IDP1
- IDP2
- SUVH2
- SUVH9
- SUVR2

Phase (iii) Chromatin alterations

- SUVH4
- IDN2
- HDA6
- JMJ14
- LDL1
- LDL2
- UBP26

Additional factors involved in methylation

- NRPB1
- NERD
- CMT2
- CMT3
- MET1
- DDM1
Fig. 5

a) ANGIOSPERMS

b) OTHER LAND PLANTS

% of total small RNAs (18-26 nt)

Small RNA size (nt)
Fig. 6

(a) V1 V2 V3 V4 V5 V6 V7

(b) C-methylation (%)

- Ginkgo biloba
- Gnetum gnemon
- Gain of:
  - RDM1
  - KTF1
  - DMS3
  - SHH1
  - SUVR2
  - NRPD4/NRPE4
- Abundant 24 nt sRNAs

- Gain of DCL2

- Secondary loss of DCL2 in Gnetales
- Low levels of CHH methylation