Exploring the evolution of the proteins of the plant nuclear envelope

AUTHORS : Axel POULET [1,2], Aline V. PROBST [2] Katja GRAUMANN[1], Christophe TATOUT [2] and David E. EVANS [1]

[1] Department of Biological and Medical Sciences, Oxford Brookes University, Oxford OX3 0BP, UK

[2] UMR CNRS 6293 INSERM U1103 Clermont Université, GReD, Aubière, France

[*] Address for correspondence David E Evans, Faculty of Health and Life Sciences, Oxford Brookes University, Headington Campus, Oxford OX3 0BP UK. deevans@brookes.ac.uk

KEY WORDS

Higher plant, Nucleus, Chromatin, LINC complex, SUN domain, KASH domain, nucleoskeleton

ABBREVIATIONS AND ACRONYMS

Arabidopsis thaliana (*A. thaliana*), Basic Local Alignment Search Tool protein (BLASTp), Crowded Nuclei (CRWN; also termed LINC for Little Nuclei and NMCP for Nuclear Matrix Constituent Protein), HMMER (Hidden Markov Model-based sequence alignment tool) Klarsicht/Anc1/Syne homology (KASH), Lamin B receptor (LBR), Lamin-Emerin-Man1 (LEM), Linker of Nucleoskeleton and Cytoskeleton (LINC), Nuclear Envelope Associated Protein (NEAP), Reads Per Kilobase of transcript per Million mapped reads (RPKM), Sad1-Unc84 (SUN), SUN interacting Nuclear Envelope Proteins (SINEs), Toll Interleukin Receptor domain KASH protein (TIK), trans-membrane (TM), Whole-genome duplication (WGD), WPP Domain Interacting Proteins (WIPs),

ABSTRACT

In this study, we explore the plasticity during evolution of proteins of the higher plant nuclear envelope (NE) from the most ancestral plant species to advanced angiosperms. The higher plant NE contains a functional Linker of Nucleoskeleton and Cytoskeleton (LINC) complex based on conserved Sad1-Unc84 (SUN) domain proteins and plant specific Klarsicht/Anc1/Syne homology (KASH) domain proteins. Recent evidence suggests the presence of a plant lamina underneath the inner membrane and various coiled-coil proteins have been hypothesised to be associated with it including Crowded Nuclei (CRWN; also termed LINC and NMCP), Nuclear Envelope Associated Protein (NEAP) protein families as well as the CRWN binding protein KAKU4. SUN domain proteins appear throughout with a key role for mid-SUN proteins suggested. Evolution of KASH domain proteins has resulted in increasing complexity, with some appearing in all species considered, while other KASH proteins are progressively gained during evolution. Failure to identify CRWN homologs in unicellular organisms included in the study and their presence in plants leads us to speculate that convergent evolution may have occurred in the formation of the lamina with each kingdom having new proteins such as the Lamin B receptor (LBR) and Lamin-Emerin-Man1 (LEM) domain proteins (animals) or NEAPs and KAKU4 (plants). Our data support a model in which increasing complexity at the nuclear envelope occurred through the plant lineage and suggest a key role for mid-SUN proteins as an early and essential component of the nuclear envelope.

INTRODUCTION

The nuclear envelope is a key component of eukaryotic cells and may be considered to be composed of three elements, the nuclear membrane, nuclear pore complexes and the nuclear lamina (Gerace and Burke, 1988; Hetzer, 2010). These structural components are essential for many processes including nuclear morphology, nuclear migration, chromatin organisation and regulation of gene expression (Graumann and Evans, 2010a). Significant progress has been made in describing novel plant nuclear envelope proteins (Parry, 2015; Tamura et al., 2015; Zhou et al., 2015a). In *Arabidopsis thaliana* (*A. thaliana*), these include components of the Linker of Nucleoskeleton and Cytoskeleton (LINC) complex for which functional data is slowly being revealed. Arabidopsis contains proteins of the inner nuclear envelope of the SUN domain family including Cter-Sad1-Unc84 (Cter-SUN) (Graumann and Evans, 2010b; Graumann et al., 2010; Oda and Fukuda, 2011) as well as mid-SUN domain proteins in which a SUN-domain homologous to that of the C-ter SUNs is located centrally within the protein (Graumann et al., 2014). It also contains proteins of the outer nuclear envelope, of the KASH domain protein family including WPP Domain Interacting Proteins [WIPs], SUN interacting Nuclear Envelope Proteins [SINEs] and *Arabidopsis thaliana* Toll

Interleukin Receptor domain KASH protein [TIK] (Zhou et al., 2012; Graumann et al., 2014; Zhou and Meier, 2014). In addition, plant proteins proposed to form the nuclear lamina - Crowded Nuclei (CRWNs; Dittmer et al., 2007; Wang et al., 2013) and CRWN-interacting proteins such as KAKU4 (Goto et al., 2014) as well as Nuclear Envelope Associated Proteins (NEAPs), which may be associated with the lamina (Pawar et al., 2016), have been described in *A. thaliana* and shown to localise to the nuclear periphery.

Sequence data now available permits comparison of components of the nuclear envelope between algae, mosses, gymnosperms and angiosperms with the components of *A. thaliana*. Functional analysis of these genes is challenging because they belong to small gene families as a consequence of gene and whole-genome duplication (WGD) creating duplicate genes and thus gene redundancy (Gaut and Ross-Ibarra, 2008; Soltis and Soltis, 2016). Whole-genome duplication (WGD) is recognised as an important event for genome evolution in animals, plants and fungi and to drive key new features, with resulting increased complexity and speciation (Soltis and Soltis, 2016). Following WGD, massive gene loss can occur restoring the diploid state for most duplicated loci while few duplicated genes remain and may provide new evolutionary innovation including structures (e.g. floral organs) and adaptations (Kellis et al., 2004). Previous analyses of plant genomes have shown that seed plants share an ancient WGD event, zeta (Jiao et al., 2011). A second WGD, epsilon, has been detected shortly before the diversification of angiosperms. These two WGDs are suggested to play a role in the origin and rapid diversification of the angiosperms (Jiao et al., 2011). Finally, the gamma WGD occurred after eudicot/monocot diversification, followed by several partial or complete duplication events.

In this study, we have selected 20 representative species based on the revised classification of eukaryotes (Adl et al., 2012), their available genome sequences and gene expression description, in order to explore the evolution of components of the plant nuclear envelope. Availability of genome sequence data for the core eudicot *Amborella trichopoda* provides an opportunity to explore the nuclear envelope of a primitive angiosperm. *Amborella*, a New Caledonian shrub, has been suggested as the sole surviving sister species of all other angiosperms and is unique in sequenced plant genomes in showing no evidence of recent, lineage-specific genome duplications (Project et al., 2013).The *Amborella* genome therefore offers an opportunity to explore the composition of an ancestral plant nuclear envelope and the effect of genomic changes after polyploidy in other angiosperms (Project et al., 2013). Finally, RNAseq data for each species were used to describe expression levels within species to establish that the genes are active and not pseudogenes and to demonstrate gene activity and when possible tissue specific expression patterns.

The aim of the work presented in this paper was to explore the evolution of nuclear envelope proteins in unicellular algae and multicellular plants and to provide evidence for the composition of the simplest functional plant LINC complex. This study provides valuable information for mutant and other functional studies by identifying potential redundancy and specialisation in nuclear envelope and lamina-like components.

MATERIAL AND METHODS

Homologous LINC complex and lamin-like protein detection

A Perl script was developed and applied to proteomic data to identify KASH domain proteins. The program tests the presence of the trans-membrane (TM) domain and four specific amino acids at the C-terminus, which are characteristic for KASH proteins. The position of the TM domain is variable and the script searches this TM domain up to 40 amino acids away from the KASH-specific C-terminal motifs detected in *A. thaliana* (either VIPT, VVPT, AVPT, PLPT, TVPT, LVPT or PPPS; Zhou et al., 2012; Graumann et al., 2014; Zhou et al., 2014). The identification of the TM domain is based on the Kyte-Doolittle method (Kyte and Doolittle, 1982). Only proteins, which possess a TM domain and the four KASH specific amino acids in the C-terminus of the protein, were selected.

For all proteins of interest a Basic Local Alignment Search Tool protein (BLASTp) was used with default parameters as well as HMMER (Hidden Markov Model-based sequence alignment tool; http//hmmer.org). The best hits were retained and used for phylogenetic analysis (Altschul et al., 1990). The proteome of each species was used as reference for the BLASTp (Figure 1), and the protein sequences of the LINC complex as well as the putative lamina of *A. thaliana* were used as queries (Supplementary Table 1). BLASTp results are given as supplementary Table 2 (mid-SUN), 3 (Cter-SUN), 4 (WIP), 5 (SINE), CRWN (6), NEAP (7) and KAKU4 (8). Reciprocal BLASTp was used to verify the relevance of all identified orthologs.

Phylogenetic reconstruction

Selected sequences were first aligned with MUSCLE, a multiple sequence alignment tool (Edgar, 2004), using default parameters. The alignment was then refined using Gblocks (Talavera and Castresana, 2007). Fast-Tree was then applied with default parameters, for the construction of the phylogenetic tree (Price et al., 2010). Fast-Tree infers approximately-maximum-likelihood phylogenetic trees from alignments. Finally, phylogenetic trees were drawn using the Interactive Tree Of Life ITOL (Letunic and Bork, 2011).

RNA sequencing data and analysis

Data used for the RNA-seq analysis was obtained from the NCBI (http://www.ncbi.nlm.nih.gov/geo/browse/) or from the Amborella Genome Database, respectively (http://amborella.huck.psu.edu/). Five different tissues (leaves, roots, flowers, flower buds, and seeds/siliques) as well as total seedling were chosen for the analysis of the expression patterns of the genes of interest (Supplementary Table 1). The expression was analysed for ten species (Supplementary Table 9). Reads from RNA-Seq libraries were mapped onto the candidate gene sequences allowing no mismatches using TOPHAT v 2.0.14 (Kim et al., 2013) with standard settings and maximum of multihits set at 1, minimum intron length set at 15 bp, and maximum intron length set as 6,000 bp. Reads were added together for each gene using HTseq-count with the overlap resolution mode set as intersection-non empty and with no strand-specific protocol (Anders et al., 2015). Transcription levels in Reads Per Kilobase of transcript per Million mapped reads (RPKM) were normalised to *AtSAND* (At2g28390; Czechowski et al., 2005 and Supplementary Table 10). *SAND* was chosen due to its constant gene expression levels across different tissues at developmental stages in *Arabidopsis thaliana* (Czechowski et al., 2005). For each species, the *SAND* homologue with the closed sequence identity to *AtSAND* was chosen. Furthermore, absolute *SAND* expression levels (in RPKM) in different species were comparable to expression of Arabidopsis *AtSAND*.

RESULTS AND DISCUSSION

In order to gain an insight into the evolutionary development of known plant nuclear envelope proteins, we reconstructed the phylogenetic distribution of the LINC complex (SUN and KASH) and plant lamina (CRWN, NEAP and KAKU4) components by exploring 20 representative species including unicellular photosynthetic algae, lycophytes, mosses, gymnosperms and angiosperms for which genome sequences and gene expression data are available (Figure 1; Supplementary Table 11).

Phylogenetic analysis of inner nuclear membrane proteins

Cter-SUN proteins

The SUNs are divided into two subfamilies according to the position of the SUN domain: the mid-SUN with a central SUN domain and the Cter-SUNs having a SUN domain at the C-terminus. The potential origin of the two classes of SUN domain proteins remains obscure and is discussed in Graumann et al., 2015. The SUNs are key members of the LINC complex and expressed in all tissues (Murphy et al., 2010; Graumann et al., 2014). Blastp

and HMMER analysis revealed thirty-three Cter-SUN proteins across all the 19 species studied, other than *Chlamydomonas reinhardtii*, where no Cter-SUN protein was detected (Figure 1). Monocots and eudicots form two paraphyletic groups, with the *Vitis vinifera* homologue showing greater similarity to the monocot Cter-SUN sequence. The *Brassicaceae* form a monophyletic group and the duplication of the Cter-SUN gene seems to have occurred late in evolution because duplicated Cter-SUNs remain grouped within a given species (Figure 2).

Expression data for the Cter-SUNs shows a similar transcript level for all the tissue analysed in different species (Supplementary Figure 1). In some cases, one of the two Cter-SUNs is more strongly expressed in the seedling (e.g.: AtSUN1 more strongly expressed than AtSUN2; OsaSUN-a more strongly expressed than OsaSUN-b) (Figure 2). *A. trichopoda* encodes only one Cter-SUN that is highly expressed in all tissues. The simplest functional LINC complex may therefore be based on a single Cter-SUN, and strengthens the suggestion that duplication of the Cter-SUN gene occurred after speciation.

One or two Cter-SUN proteins were identified in most plants and the moss, although four close homologues were identified for the club moss *Selageinella molendorfii*. In *A. thaliana*, SUN1 and SUN2 share almost the same activity and localisation (Graumann et al., 2010). This is in contrast to mammals, where five Cter-SUN orthologues have clearly differentiated functions. It appears that the gene duplication resulting in these orthologues occurred earlier in the evolution of mammals. One likely consequence is the lack of specificity of function of plant Cter-SUN homologues; for example, a disruption of a single SUN gene results in an infertility phenotype in animals (Ding et al., 2007), but in *A. thaliana*, a single Cter-SUN deletion does not affect meiosis or fertility whereas the double mutant *atsun1 atsun2* impacts fertility and cell division (Varas et al., 2015). This suggests a significant redundancy in Cter-SUN function in plants and that double knock-out or knock-down mutants are required for recognisable phenotypes to be obtained.

mid-SUN proteins

All the species considered contain at least a mid-SUN protein and overall fifty mid-SUN homologues were identified during our bioinformatic screen. The mid-SUN angiosperm homologues are clustered in two groups, SUN3/SUN4 and SUN5. In each mid-SUN homologous group, the basal angiosperm, monocots and eudicots form monophyletic groups. This suggests that mid-SUN (3, 4) gene duplication occurred after speciation between angiosperms and gymnosperms (Figure 3). In all tissue analysed the *SUN3/SUN4* group tends to be more ubiquitously and highly expressed than the *SUN5* group, this is also

true for the *A. trichopoda* homologues (Supplementary Figure 1). It has been suggested that *AtSUN5* has a meiotic function (Graumann et al., 2014) and this is also true for maize with *ZmaSUN5* (Murphy et al., 2010), although while the double mutants of SUN3, SUN4 and SUN5 are viable, a *sun3 sun4 sun5* triple mutant is lethal (Graumann et al., 2014). *A. trichopoda* has two mid-SUN proteins, one SUN3/SUN4 homologue and a SUN5 homologue. This suggests that the simplest LINC complex has two mid-SUNs each with a specific or partially overlapping function.

In summary, the majority of the 20 species possess at least one mid-SUN and one Cter-SUN protein except for *Chlamydomonas reinhardtii,* which has only one mid-SUN protein. Interestingly, in common with Cter SUNs, the club moss Selaginella has the highest number of mid-Sun protein homologues (six). These results are in good agreement with previous studies that have highlighted the conservation of both Cter- and mid-SUN proteins in most eukaryotes (Murphy et al., 2010; Graumann et al., 2014) and suggest the that the LINC complex was present in the Last Evolutionary Common Ancestor (LECA; Koreny and Field, 2016). Mid-SUN homologues and Cter-SUN proteins were detected in the unicellular algae examined, suggesting that SUN emergence pre-dates the evolution of multicellularity. The evolutionary relationship between Cter-SUN and mid-SUN proteins has yet to be described. This study suggests that SUN domain proteins may also be among the earliest evolving components of the plant nuclear envelope and that mid-SUN proteins may have significance nuclear function in the absence of Cter-SUN.

KASH protein homologues

KASH proteins are diverse in sequence and structure (Zhou and Meier, 2014) but possess a conserved C-terminal region with a TM domain and a conserved motif of four amino acids at the extreme C-terminus. For the detection of the KASH domain protein homologues, two strategies were used. The first was a BLASTp analysis based on known *A. thaliana* KASH domain proteins (Supplementary Table 1). This analysis permitted detection of KASH protein homologues in all the organisms studied, except for the unicellular algae where no KASH protein was detected (Supplementary Table 12). Using this method, 32 SINE homologues were found, whereas WIP and potential TIK proteins [or TIK-like proteins] (Supplementary Table 12) were much less common and were found mainly in eudicots. An exception is *Brassicaceae*, where several potential WIP (3 in *A. lyrata*, 4 in *Brassica rapa*) and TIK (2 in *A. lyrata* and 1 in *B. rapa*) homologues were identified, these were also detected in *Glycine max* (2 WIP, 1 TIK), *Prunus persica* (1 TIK), *Carica papaya* (1 WIP), *Musa acuminata* (1 WIP), *A. trichopoda* (1 TIK) and the gymnosperm *Picea abies* (1 TIK) (Supplementary Table

12). To expand the data collected by Blastp, a script was developed to detect proteins with the TM domain and C-terminal motif.

All the identified plant KASH domain proteins have been divided into three groups: SINEs, WIPs and TIK (Zhou and Meier, 2014). Six KASH protein clusters were revealed (Supplementary Figure 2). One includes WIP proteins detected in the monocotyledons and the basal angiosperms (Supplementary Table 12), as well as seven new putative WIP proteins to those detected previously by BLASTp. For SINE proteins, three clusters were detected, for SINE1/2, SINE3 and SINE4 adding respectively two, six and twelve SINE proteins to those already identified. The high number of proteins in the SINE3 and SINE4 cluster found only by the script was due to weak conservation of these proteins. One much smaller cluster includes the TIK-like proteins. Only four putative homologues were added but these were shown subsequently to lack either the TIR domain or the C-terminal TM domain, therefore suggesting that the TIK protein (Graumann et al. 2014) may be unique to *A. thaliana*. An additional cluster (other) had low sequence similarity and was not included subsequently. The three WIP proteins in *A. thaliana* show previously described properties (Xu et al., 2007; Zhao et al., 2008) of a cytoplasmic domain at the N-terminus, with AtWIP1 and AtWIP2 having three coiled-coil domains but AtWIP3 only one. The C-terminal region is well conserved and the coiled-coil domains align, with all proteins detected as homologues having a C-terminal predicted TM domain and KASH motif, except AlyWIP1, which lacks homology at the C-terminal region but is well conserved at the N-terminus.

All SINEs have a typical KASH TM domain and C-terminal amino acid motif (Zhou et al., 2014). The SINE gene family comprises four genes in *A. thaliana*, with similarity between AtSINE1 and AtSINE2, characterised by an Armadillo repeat domain near the N-terminus; and between AtSINE3 and AtSINE4. The Perl script added only two sequences in the SINE1/SINE2 group while it added 14 proteins to the SINE3/SINE4 cluster. In this case the Blastp approach was less efficient than the Perl script because of the absence of well-conserved domains in the N-terminus. After removal of sequences with the lowest similarity or without the conserved domain, SINE1/SINE2 proteins are present in all species except the unicellular algae and club moss, while SINE3/SINE4 were absent (Figure 5). In summary, the SINE1/SINE2 cluster and WIP proteins are detected in basal angiosperms whereas the TIK protein is detected only in *A. thaliana*.

Phylogenetic analysis of the outer nuclear membrane proteins

WIP proteins

The WIP protein family was the first KASH family detected in *A. thaliana*, (Zhou et al., 2012). WIP proteins were not detected in unicellular algae, moss, club moss or gymnosperms; suggesting that they are angiosperm specific proteins. One WIP homologue was detected

for *A. trichopoda*. The monocots form a monophyletic group, with one protein for rice , two and three for maize and *Musa acuminata* suggesting gene duplication (Figure 4). The eudicots form a paraphyletic group because the WIP homologue of *Nelumbo nucifera* differs from, and is positioned outside, the WIPs of eudicots. The *Brassicaceae* on the other hand, form a monophyletic group (Figure 4). This suggests that an ancestral duplication in the *Brassicaceae* ancestor gave rise to WIP1/WIP2 and WIP3, and then WIP1 and WIP2 resulted from a more recent gene duplication. All three genes are expressed in all the tissues analysed. In *A. thaliana AtWIP3* transcripts are more abundant than *AtWIP1* and *AtWIP2* in all tissues. This may be due to redundancy in AtWIP1 and AtWIP2 function, and in *A. trichopoda*, the WIP homologue is highly expressed (Supplementary Figure 1).

SINE proteins

SINEs in *A. thaliana* (Zhou et al., 2014) comprise two groups, SINE1/SINE2 and SINE3/SINE4. AtSINE1 is more expressed in guard cells, and its armadillo domain forms F-actin-associated fibres involved in nuclear positioning while AtSINE2 is suggested to be involved in the immunity response of leaves (Zhou et al., 2014). No expression and activity data was available for AtSINE3 and AtSINE4.

SINE1/SINE2 proteins were not found in unicellular algae and in club moss, but in contrast to WIPs, two and three SINE homologues were found in moss and gymnosperms, respectively (Figure 5). The angiosperms form a monophyletic group and one SINE1/SINE2 homologue was detected for *A. trichopoda* and positioned at the base of the angiosperm group (Figure 5). The phylogenetic analysis of SINE3 and SINE4 is not possible due to the low similarity between sequences and a lack of conserved domains. Although SINE3 and SINE4 are detected in the *Brassicaceae* group, the other sequences are divergent. In the monocots, two protein homologues were detected for *Musa acuminata*, *Oryza sativa* and *Zea mays*. However, the phylogeny suggests the presence of recent gene duplication in *Musa acuminata* (Figure 5). In contrast, the gene duplication between the two other monocots seems to have occurred before their speciation. All the eudicots possess at least one SINE1/SINE2 homologue. Four homologues that group together were found in *Glycine max*, suggesting a recent gene duplication. As for WIPs, *Brassicaceae* proteins cluster together, and one group of homologues is detected for each of SINE1 and SINE2. The organisation between the two groups suggests a gene duplication to form SINE1 and SINE2.

In *A. thaliana, AtSINE1* and *AtSINE2* are expressed at the same level in all tissues, but at a higher level than *AtSINE3* and *AtSINE4*. However, SINE1/SINE2 homologues in most other species show the lowest level of expression of all KASH proteins for all tissues analysed expect for maize, rice and *A. trichopoda* (Supplementary Figure 1). In these

species WIP and SINE expression is at the same level for all tissues. In *A. thaliana*, *AtWIPs* are more highly expressed than *AtSINE*.

Phylogenetic analysis of the putative nuclear lamina and nuclear-envelope associated proteins

Proteins of the lamin family are restricted to animals (Cavalier-Smith, 2010). However, to date, three protein families have been suggested to be components of the putative lamina in *A. thaliana*, CRWN (Dittmer et al., 2007; Wang et al., 2013), KAKU4, (Goto et al., 2014) and a novel nuclear envelope associated protein family, NEAP (Pawar et al., 2016).

CRWN proteins

The CRWN gene family is imported into the nucleus through an NLS and extensive coiled-coil domains reminiscent of the animal lamins are hypothesised to allow polymerisation of the protein to form the plant lamina. Fifty CRWN proteins were detected by BLASTp and pHMMER in all multicellular plants but are absent from unicellular algae. In most species two homologues were detected for each species. Two clusters of CRWN proteins were defined in a previous publication (Ciska and Moreno Diaz de la Espina, 2013) and were also identified here as two main phylogenetic groups: CRWN1/CRWN2/CRWN3 and CRWN4. The clusters of CRWN4 homologues constitute monophyletic groups and only one protein was found for all species except for *Glycine max* (Figure 6). Gymnosperm homologues seem to have only the CRWN4 lineage while *A. lyrata* has lost the CRWN4 lineage meaning that some functional redundancy exists between the two monophyletic groups.

For the second group made up of the homologues of the three other CRWN proteins, the same organisation was found and only one homologue in *A. trichopoda* was detected (Figure 6). In the monocot group, only *Musa acuminata* possesses three homologues, the other monocots possessing only one (Figure 6). In the eudicot group, two clusters can be distinguished: one for the homologues of AtCRWN1 and the other for AtCRWN2/AtCRWN3. This reveals a gene duplication, which occurred after the speciation creating monocots and eudicots. The other duplication, which gave rise to CRWN2 and CRWN3, occurred after *Brassicaceae* speciation and formed a monophyletic group. The genes belonging to the cluster CRWN1/CRWN2/CRWN3 show higher expression in comparison to CRWN4. Other than in the *Brassicaceae*, CRWN2 is less expressed than CRWN1 and CRWN3 for all the tissues analysed. Surprisingly, no lamin-like proteins were detected in the chlorophyte unicellular algae. Previous studies have shown the presence of other lamin-like proteins in unicells like NE81 and NUP1 (DuBois et al., 2012; Krüger et al., 2012). It is likely that several

proteins have evolved in different systems to fulfil a similar role and this would reward further study.

NEAP proteins

The NEAP proteins are characterised by a TM region at the C-terminus, a functional NLS and extensive coiled-coil domains (Pawar et al., 2016). NEAP1, NEAP2, and NEAP3 were identified in gymnosperms and angiosperms and 28 proteins were detected while NEAPs are absent from the more ancestral species moss, club moss and unicellular algae (Supplementary table 11). The monocots form a monophyletic group with two potential specific gene duplications for *Musa acuminata* and *Zea mays* (Figure 7). As for monocots, the eudicots form a monophyletic group (Figure 7), and the gene duplication seems specific to species. So the three NEAP genes in *Brassicaceae* appear to result from a duplication event during the speciation of *Brassicaceae*. The single NEAP gene in *A. trichopoda* is expressed at very high level. Lack of expression of *AtNEAP4* and absence of protein homologues in other species imply that it is a pseudogene. The other *NEAP* genes are expressed in seedlings and in other tissues but at a low level (Supplementary Figure 1).

KAKU4 proteins

KAKU4 homologues are only detected in angiosperms. Only one KAKU4 homologue is detected in each species except for *Glycine max* and *Brassica rapa*. KAKU4 is therefore a recent addition specific to angiosperms and as it interacts with CRWN1 and CRWN4, it could link CRWN proteins to other components at the nuclear periphery. Analysis of KAKU4 phylogeny reveals two monophyletic groups, for the monocot and eudicot homologues (Figure 8). The protein was not detected in basal angiosperms, gymnosperms, moss, club moss and unicellular algae. Either KAKU4 is a protein with specific function in angiosperms, or was not detected due to a high variability between species. KAKU4 homologues are expressed to comparable levels in all tissues (Supplementary Figure 1).

DISCUSSION

The results presented reveal functional conservation of the proteins of the plant nuclear envelope with those of other kingdoms, but surprising diversity in protein sequence. Table 1 summarises the occurrence of each of the components in the study, together with their function. SUN domain proteins, lamina component CRWN and the KASH domain proteins SINE1-2 involved in actin binding are present before the Zeta WGD (though C-ter SUNs and CRWN are absent from the Chlamydomonas); putative NE- anchored lamina components NEAPs are first found after the Zeta WGD. Binding of RanGAP to the NE by the KASH

proteins designated WIP originates with the gamma WGD of the angiosperms; the mechanism of anchorage of RanGAP in gymnosperms and mosses therefore warrants further study. CRWN interacting KAKU4 and KASH protein TIK appear to be of later origin and were only detected in the *Brassicas*, suggesting specialist functions.

Data from *A. trichopoda* suggests a minimal angiosperm LINC complex, with two KASH domain proteins (one WIP and one SINE), three SUN domain proteins (one Cter-SUN and two mid-SUNs) and putative lamina constituents (two CRWNs) together with one NEAP. The moss *P. patens* has four SUNs, two KASH and two putative lamina constituents. The gymnosperm *P. abies* has three SUNs, three KASH (SINEs) and putative lamina components (two CRWN) together with two NEAPs (Figure 1). Evolution of SUN, KASH and lamina constituents appears to have accompanied WGD and partial genome duplication events and to have resulted in a range of homologues during angiosperm speciation. The results presented also indicate a plant nuclear envelope which has developed significant complexity and redundancy through gene duplication explaining the need for multiple knock-out mutants; for instance the double mutant *atsun1 atsun2* (Zhou et al., 2012) or the quintuple mutant *wifi* (*atwip1 atwip2 atwip3 atwit1 atwit2*) (Zhou et al., 2015b) before strong phenotypes are observed.

The data also suggests evolution from an ancestral LINC complex, in which SUN domain proteins are multifunctional, to a more multifaceted LINC complex containing an increasing number of KASH and lamin-like proteins. A key role for mid-SUN proteins is suggested. This is commensurate with the demonstration that SUNs play a fundamental role in chromatin interaction with the nuclear envelope during its reformation in plant mitosis (Graumann and Evans, 2011) and in telomere attachment in meiosis (Varas et al., 2013). KASH domain proteins appear to be evolving, with SINEs preceding WIPs, with TIK only identified in Arabidopsis. It is suggested that increasing specialisation accompanies the acquisition of additional KASH homologues and that specific functions of the later evolving proteins (for instance RanGTP anchorage and nuclear movement in the pollen tube) are undertaken by other nuclear envelope components in their absence. A similar pattern of evolution of KASH proteins is suggested in ophisthokonts; Zhou et al., (2014) commenting on novel plant KASH proteins noted that while some are highly conserved (e.g. Nesprin 1 and 2, ANC-1 and MSP-300) others are restricted in distribution (e.g. Klarsicht homologues to insects and KDP-1 to nematodes) suggesting origins after SUN domain proteins and rapid evolution linked to diversifying function. Finally, higher plants have evolved a lamina-like structure based, like animal cells, on coiled-coil proteins. This appears to have arisen with the CRWN proteins present in mosses and clubmosses (lycophytes) and with KAKU4 arising later. These data are consistent with previous reports suggesting that SUN domain proteins were the first

nuclear envelope proteins linking chromatin to the nuclear envelope (Cavalier-Smith, 2010), predating the evolution of lamins. Indeed lamins are prone to rapid evolution as they interact with fewer partners than components of the LINC complex (Koreny et al 2016). One of the striking results of our analysis is the absence of CRWN in unicellular species despite the fact that the lamina is involved in basic function such as nuclear morphology and chromatin organisation which are both important for the regulation of gene expression. Although we cannot exclude if lamins and CRWN are subjected to fast evolution leading to unsuccessful recovery of homologs by Blast and HMMMER analyses, it is tempting to speculate that convergent evolution occurred in animals and plants, with increasing functionality and complexity through the introduction of LBR and LEM proteins in animal and KAKU4 in plants. Similar observations were recently presented for the PRC1 polycomb group complex which is a conserved function but with poor sequence homology despite the presence of the conserved RING-domain, again suggesting convergent evolution between plant and animals. Interestingly, the PRC1 complex is involved in the regulation of gene expression through the binding of trimethylated Histone H3 at lysine 27 (H3K27me3) a well-known repressive epigenetic mark also enriched in Lamina-Associated Domains (LADs) (Bickmore and van Steensel, 2013). Exploring the possible connection between the nuclear envelope components and the PRC1 repressive complex will lead to a better understanding of the functions of the nuclear envelope in the regulation of chromatin organisation and gene expression.

ACKNOWLEDGEMENTS

The work was supported by the CNRS, INSERM, Blaise Pascal and Oxford Brookes Universities. KG is a Leverhulme Trust Early Career Fellow.

462 Table 1

	First appearance	WGD	Function	Location	Reference
C-ter SUN	Moss		LINC complex component; binds KASH; Nuclear shape and size; meiosis	INM	Graumann et al., 2010; Oda and Fukuda, 2011
Mid-SUN	Alga		LINC complex component; binds KASH; Nuclear shape and size; fertility	INM and ER	Graumann et al., 2014
CRWN	Moss		nucleoskeleton; nuclear size and shape; heterochromatin organisation	nuclear periphery and nucleoplasm	Dittmer et al., 2007; Wang et al., 2013
SINE1-2	Moss		LINC complex component; KASH; interacts with actin cytoskeleton; nuclear positioning in guard cells (SINE1); innate immunity response (SINE2)	ONM	Zhou et al., 2012, 2014
NEAP	Basal angiosperm, Gymnosperm	Zeta	NE anchor, SUN binding, chromatin interactor; root growth; nuclear morphology	INM	Pawar, 2015
WIP	Basal angiosperm	Epsilon	SUN binding; anchors RanGAP to NE; nuclear morphology; pollen tube termination; nuclear movement	ONM; RanGAP anchorage	Xu et al., 2007; Zhao et al., 2008
KAKU 4	Monocot	Epsilon	CRWN binding; nuclear size and shape	Nucleoskeleton	Goto et al., 2014
SINE 3-4	Eudicot (Brassicas)	Gamma	LINC complex component; KASH; SUN binding	ONM	Zhou et al., 2014
TIK	Eudicot (Arabidopsis)	Gamma	SUN binding; nuclear size and shape; root growth	ONM	Graumann et al., 2014

463

464 Table 1: protein classes and their origins and function as derived in this study.

465

466

467

REFERENCES

Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., et al. (2012). The Revised Classification of Eukaryotes. J. Eukaryot. Microbiol. *59*, 429–514.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Bickmore, W.A., and van Steensel, B. (2013). Genome Architecture: Domain Organization of Interphase Chromosomes. Cell *152*, 1270–1284.

Cavalier-Smith T (2010) Kingdoms protozoa and chromista and the eozoan root of the eukaryotic tree. Biology Letters 6, 342-345

Ciska, M., and Moreno Diaz de la Espina, S. (2013). NMCP/LINC proteins: putative lamin analogs in plants? Plant Signal. Behav. *8*, e26669.

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M.K., and Scheible, W.-R. (2005). Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis. Plant Physiol. *139*, 5–17.

Ding, X., Xu, R., Yu, J., Xu, T., Zhuang, Y., and Han, M. (2007). SUN1 Is Required for Telomere Attachment to Nuclear Envelope and Gametogenesis in Mice. Dev. Cell *12*, 863–872.

Dittmer, T.A., Stacey, N.J., Sugimoto-Shirasu, K., and Richards, E.J. (2007). LITTLE NUCLEI Genes Affecting Nuclear Morphology in Arabidopsis thaliana. Plant Cell *19*, 2793–2803.

DuBois, K.N., Alsford, S., Holden, J.M., Buisson, J., Swiderski, M., Bart, J.-M., Ratushny, A.V., Wan, Y., Bastin, P., Barry, J.D., et al. (2012). NUP-1 Is a Large Coiled-Coil Nucleoskeletal Protein in Trypanosomes with Lamin-Like Functions. PLoS Biol. *10*.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Gaut, B.S., and Ross-Ibarra, J. (2008). Selection on Major Components of Angiosperm Genomes. Science *320*, 484–486.

Gerace, L., and Burke, B. (1988). Functional organization of the nuclear envelope. Annu. Rev. Cell Biol. *4*, 335–374.

Goto, C., Tamura, K., Fukao, Y., Shimada, T., and Hara-Nishimura, I. (2014). The Novel Nuclear Envelope Protein KAKU4 Modulates Nuclear Morphology in Arabidopsis. Plant Cell *26*, 2143–2155.

Graumann, K., and Evans, D.E. (2010a). The plant nuclear envelope in focus. Biochem. Soc. Trans. *38*, 307–311.

Graumann, K., and Evans, D.E. (2010b). Plant SUN domain proteins: Components of putative plant LINC complexes? Plant Signal. Behav. *5*, 154–156.

507 Graumann, K., Runions, J., and Evans, D.E. (2010). Characterization of SUN-domain
508 proteins at the higher plant nuclear envelope. Plant J. *61*, 134–144.

509 Graumann, K., Vanrobays, E., Tutois, S., Probst, A.V., Evans, D.E., and Tatout, C. (2014).
510 Characterization of two distinct subfamilies of SUN-domain proteins in Arabidopsis and their
511 interactions with the novel KASH-domain protein AtTIK. J. Exp. Bot. eru368.

512 Hetzer, MW., The Nuclear Envelope, Cold Spring Harbour Perspectives in Biology, 2010;2
513 a000539

514 Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E.,
515 Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed
516 plants and angiosperms. Nature *473*, 97–100.

517 Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient
518 genome duplication in the yeast Saccharomyces cerevisiae. Nature *428*, 617–624.

519 Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013).
520 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and
521 gene fusions. Genome Biol. *14*, R36.

522 Koreny, L., and Field, M.C. (2016). Ancient eukaryotic origin and evolutionary plasticity of
523 nuclear lamina. Genome Biol. Evol.

524 Krüger, A., Batsios, P., Baumann, O., Luckert, E., Schwarz, H., Stick, R., Meyer, I., and Gräf,
525 R. (2012). Characterization of NE81, the first lamin-like nucleoskeleton protein in a
526 unicellular organism. Mol. Biol. Cell *23*, 360–370.

527 Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character
528 of a protein. J. Mol. Biol. *157*, 105–132.

529 Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of
530 phylogenetic trees made easy. Nucleic Acids Res. *39*, W475–W478.

531 Murphy, S.P., Simmons, C.R., and Bass, H.W. (2010). Structure and expression of the
532 maize (Zea mays L.) SUN-domain protein gene family: evidence for the existence of two
533 divergent classes of SUN proteins in plants. BMC Plant Biol. *10*, 269.

534 Oda, Y., and Fukuda, H. (2011). Dynamics of Arabidopsis SUN proteins during mitosis and
535 their involvement in nuclear shaping. Plant J. *66*, 629–641.

536 Parry, G. (2015). The plant nuclear envelope and regulation of gene expression. J. Exp. Bot.
537 *66*, 1673–1685.

538 Pawar, V. (2016). A novel family of plant nuclear envelope associated proteins. J. Exp. Bot.
539 accepted.

540 Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-
541 Likelihood Trees for Large Alignments. PLoS ONE *5*, e9490.

542 Project, A.G., Albert, V.A., Barbazuk, W.B., dePamphilis, C.W., Der, J.P., Leebens-Mack, J.,
543 Ma, H., Palmer, J.D., Rounsley, S., Sankoff, D., et al. (2013). The Amborella Genome and
544 the Evolution of Flowering Plants. Science *342*, 1241089.

Soltis, P.S., and Soltis, D.E. (2016). Ancient WGD events as drivers of key innovations in angiosperms. Curr. Opin. Plant Biol. *30*, 159–165.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564–577.

Tamura, K., Goto, C., and Hara-Nishimura, I. (2015). Recent advances in understanding plant nuclear envelope proteins involved in nuclear morphology. J. Exp. Bot. *66*, 1641–1647.

Varas, J., Graumann, K., Osman, K., Pradillo, M., Evans, D.E., Santos, J.L., and Armstrong, S.J. (2015). Absence of SUN1 and SUN2 proteins in Arabidopsis thaliana leads to a delay in meiotic progression and defects in synapsis and recombination. Plant J. *81*, 329–346.

Wang, H., Dittmer, T.A., and Richards, E.J. (2013). Arabidopsis CROWDED NUCLEI (CRWN) proteins are required for nuclear size control and heterochromatin organization. BMC Plant Biol. *13*, 200.

Xu, X.M., T. Meulia, and I. Meier. 2007. Anchorage of plant RanGAP to the nuclear envelope involves novel nuclear-pore-associated proteins. Curr. Biol. 17:1157-1163.

Zhao, Q., J. Brkljacic, and I. Meier. 2008. Two distinct interacting classes of nuclear envelope-associated coiled-coil proteins are required for the tissue-specific nuclear envelope targeting of Arabidopsis RanGAP. Plant Cell. 20:1639-1651.

Zhou, X., and Meier, I. (2014). Efficient plant male fertility depends on vegetative nuclear movement mediated by two families of plant outer nuclear membrane proteins. Proc. Natl. Acad. Sci. U. S. A. *111*, 11900–11905.

Zhou, X., Graumann, K., Evans, D.E., and Meier, I. (2012). Novel plant SUN–KASH bridges are involved in RanGAP anchoring and nuclear shape determination. J. Cell Biol. *196*, 203–211.

Zhou, X., Graumann, K., Wirthmueller, L., Jones, J.D.G., and Meier, I. (2014). Identification of unique SUN-interacting nuclear envelope proteins with diverse functions in plants. J. Cell Biol. *205*, 677–692.

Zhou, X., Graumann, K., and Meier, I. (2015a). The plant nuclear envelope as a multifunctional platform LINCed by SUN and KASH. J. Exp. Bot. *66*, 1649–1659.

Zhou, X., Groves, N.R., and Meier, I. (2015b). Plant nuclear shape is independently determined by the SUN-WIP-WIT2-myosin XI-i complex and CRWN1. Nucl. Austin Tex *6*, 144–153.

FIGURE LEGENDS.

Figure 1: Distribution of components of plant nuclear envelope in the plant kingdom.

A) Selected plant lineages used in this study from left to right: Unicells Algae (pink), Moss and Club Moss (red), Gymnosperm (orange), Basal Angisoperms (yellow), Monocots (green)

584 and Eudicots (blue). zeta epsilon and gamma WGDs are indicated as arrow heads
585 respectively in black, grey and purple. **B)** Distribution of the 9 protein families (rows) in the
586 20 species (columns). Absence (0) of a given protein is highlighted in light orange.
587

588 **Figure 2: Phylogenetic tree of Cter-SUN proteins and gene expression levels.**
589 **Left:** maximum likelihood tree of Cter-SUN protein homologues constructed from an
590 alignment. Bootstrap values are presented. The colour of the label shows the lineage of the
591 plant. The gene label is constructed with the three letters from the species name
592 (supplementary Table 4) and the gene name of the *A. thaliana* homologues. **Right:** red bar
593 represents the value of the transcription level in seedlings expressed in RPKM, except for
594 species indicated by *, the RNA-seq data was obtained from leaf tissue (Supplementary
595 Table 2).
596

597 **Figure 3: Phylogenetic tree of mid-SUN proteins.**
598 Legend as Figure 2.
599

600 **Figure 4: Phylogenetic tree of WIP proteins.**
601 Legend as Figure 2.
602

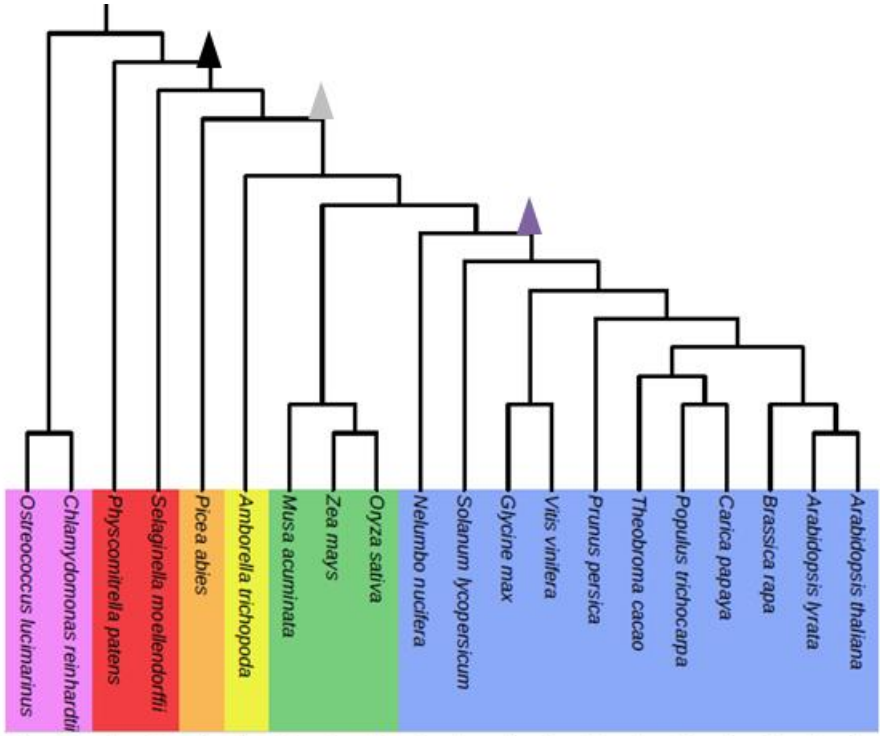
603 **Figure 5: Phylogenetic tree of SINE1, SINE2 homologues proteins.**
604 Legend as Figure 2.
605

606 **Figure 6: Phylogenetic tree of CRWN proteins.**
607 Legend as Figure 2.
608

609 **Figure 7: Phylogenetic tree of NEAP proteins.**
610 Legend as Figure 2.
611

612 **Figure 8: Phylogenetic tree of KAKU4 proteins.**
613 Legend as Figure 2.

A



B

Mid-SUN	1	1	2	6	1	2	2	2	2	2	1	6	2	2	2	3	3	4	3	3
Cter-SUN	1	**0**	2	4	2	1	3	2	2	1	1	2	1	1	1	2	1	2	3	2
CRWN	**0**	**0**	2	1	2	2	4	2	2	3	3	7	3	3	3	3	3	5	2	4
SINE1-2	**0**	**0**	2	**0**	3	1	2	2	2	2	1	4	1	1	1	2	1	3	3	2
NEAP	**0**	**0**	**0**	**0**	2	1	2	2	1	2	1	2	2	2	1	2	1	4	3	4
WIP	**0**	**0**	**0**	**0**	**0**	1	1	2	1	1	1	3	1	1	1	2	1	3	3	3
KAKU4	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**	1	1	2	1	1	1	1	1	2	1	1
SINE3-4	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	3	2	2
TIK	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1

Figure 1: Distribution of components of plant nuclear envelope in the plant kingdom.

A) Selected plant lineages used in this study from left to right: Unicells Algae (pink), Moss and Club Moss (red), Gymnosperm (orange), Basal Angioperms (yellow), Monocots (green) and Eudicots (blue). zeta epsilon and gamma WGDs are indicated as arrow heads respectively in black, grey and purple. **B)** Distribution of the 9 protein families (rows) in the 20 species (columns). Absence (0) of a given protein is highlighted in light orange.

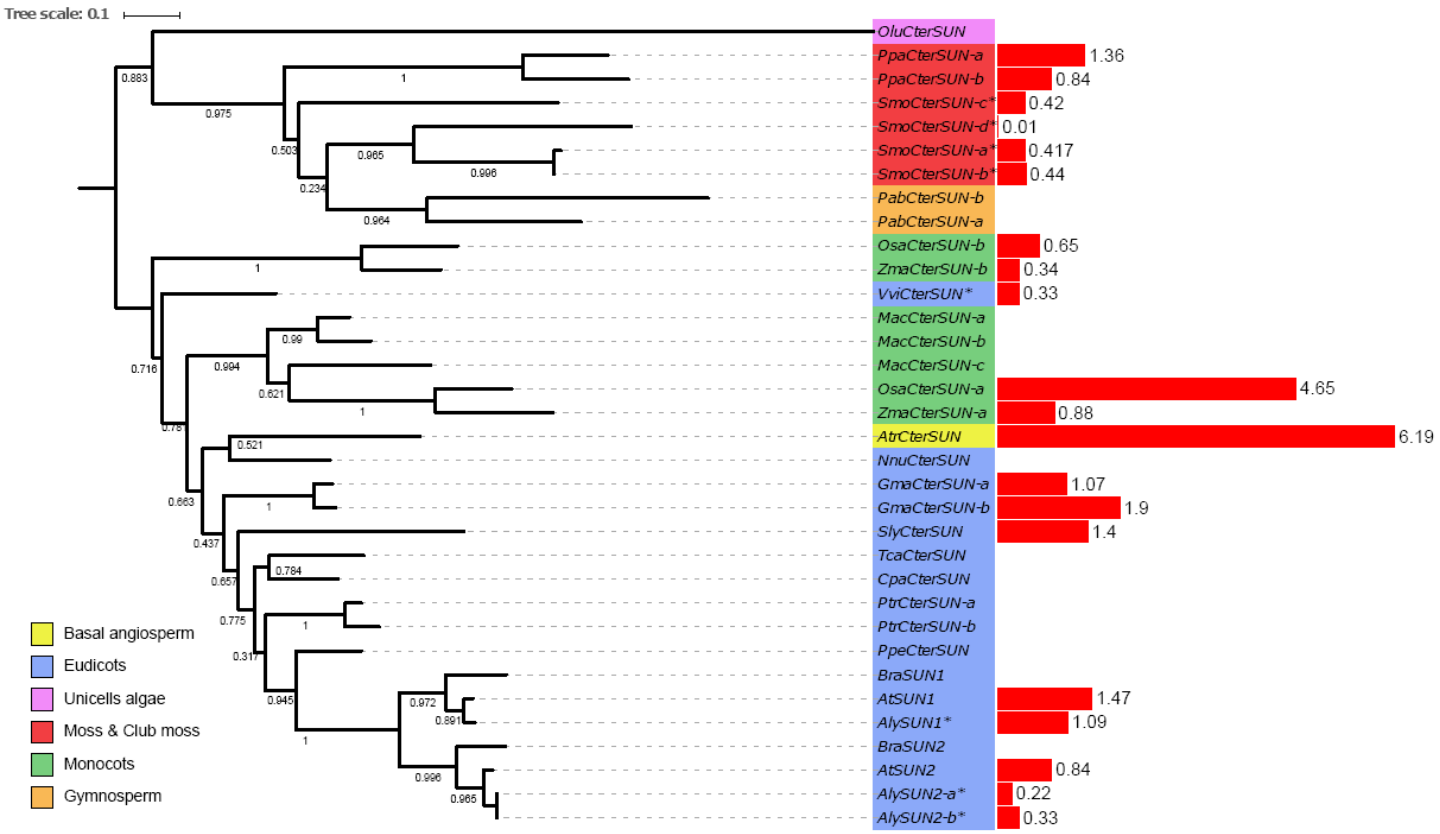


Figure 2: Phylogenetic tree of Cter-SUN proteins and gene expression levels.

Left: maximum likelihood tree of Cter-SUN protein homologues constructed from an alignment. Bootstrap values are presented. The colour of the label shows the lineage of the plant. The gene label is constructed with the three letters from the species name (supplementary Table 4) and the gene name of the *A. thaliana* homologues. **Right:** red bar represents the value of the transcription level in seedlings expressed in RPKM, except for species indicated by *, the RNA-seq data was obtained from leaf tissue (Supplementary Table 2).

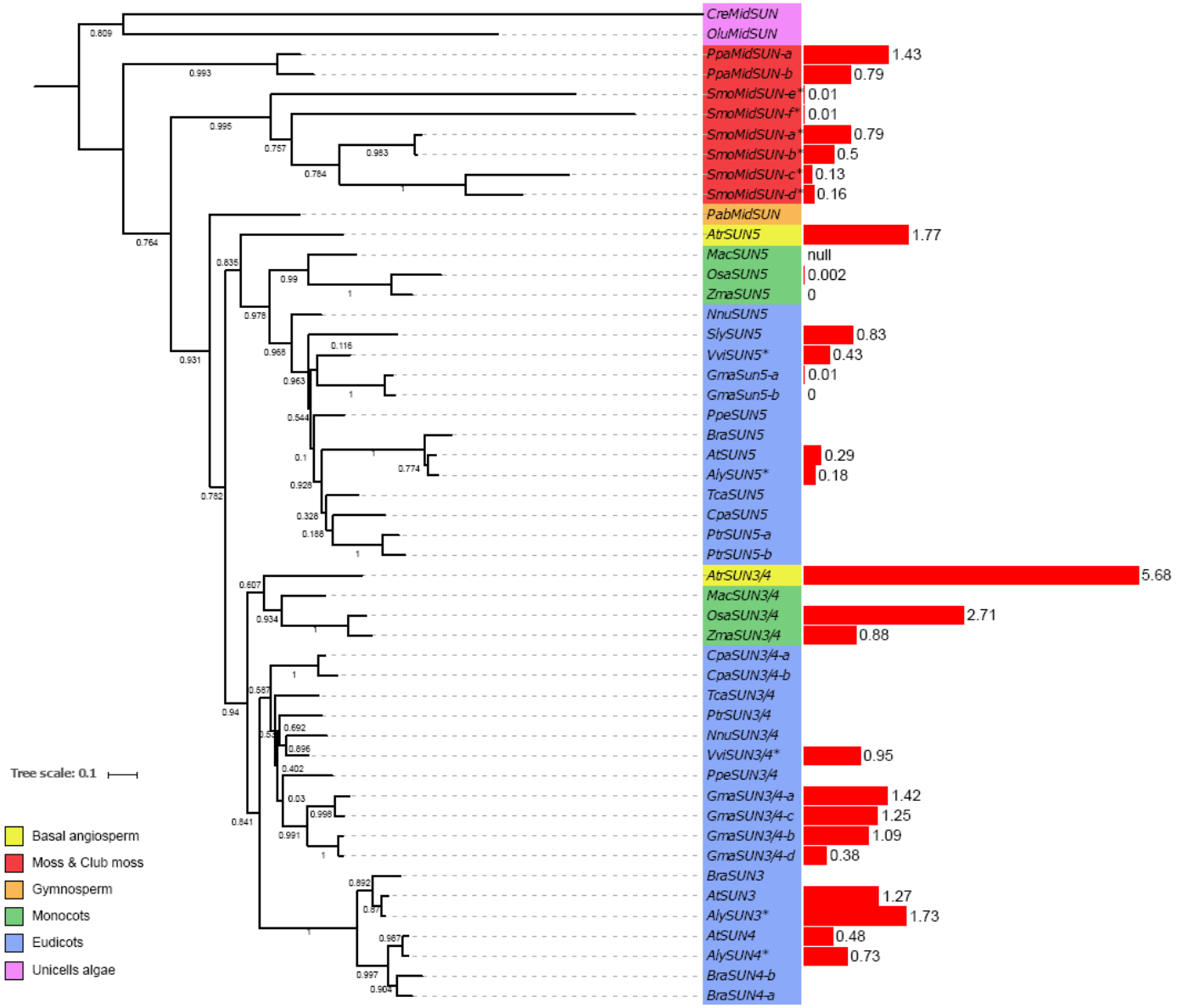


Figure 3: Phylogenetic tree of Mid-SUN proteins.
Legend as in Figure 2.

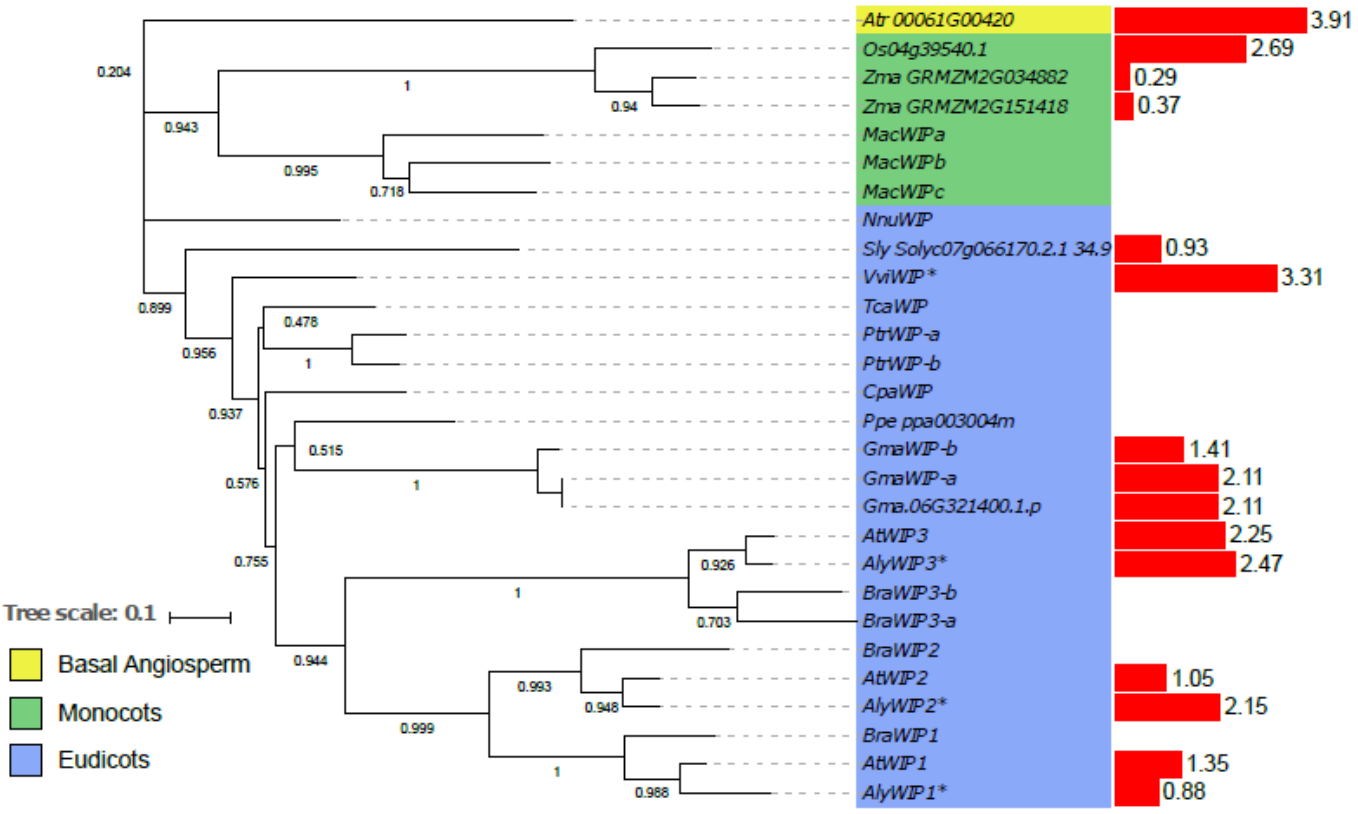


Figure 4: Phylogenetic tree of WIP proteins.
Legend as in Figure 2.

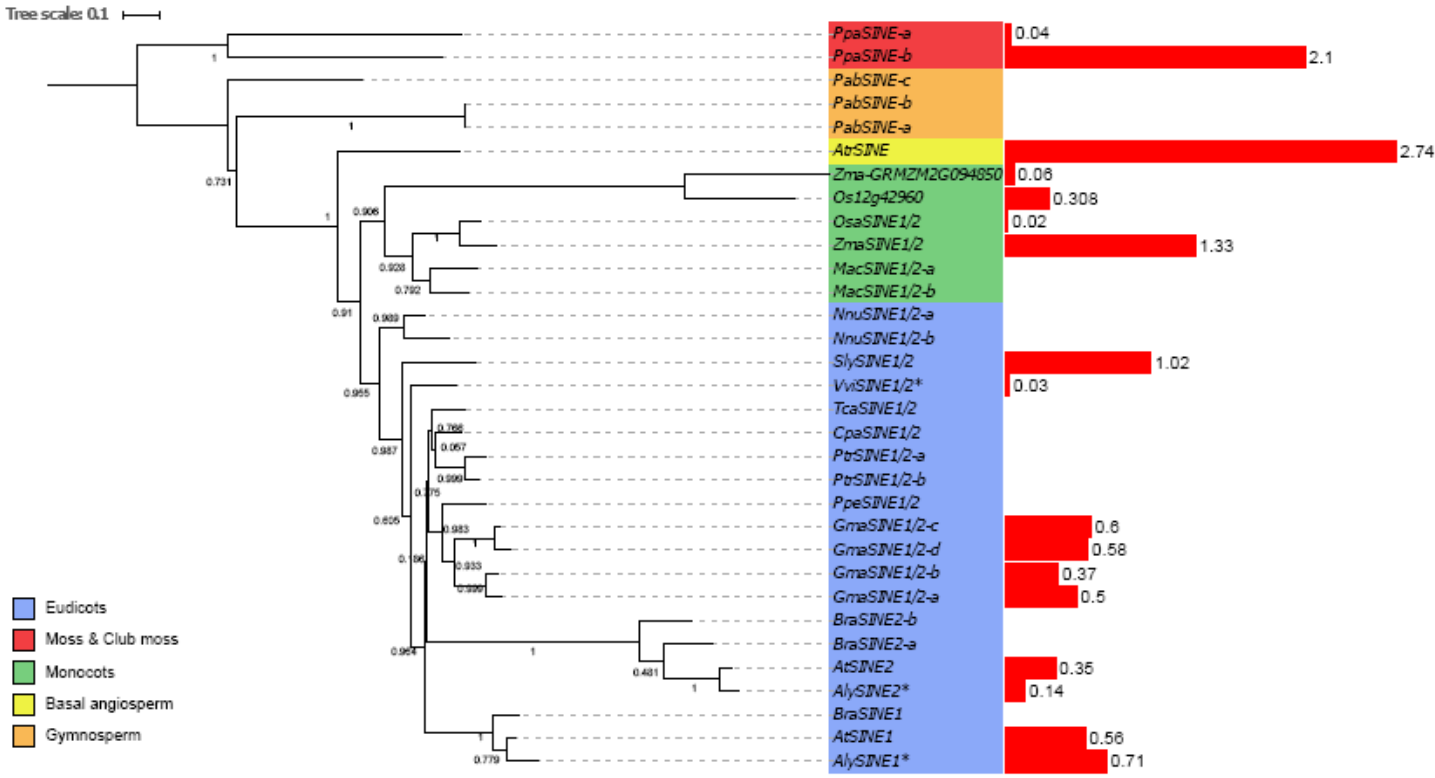


Figure 5: Phylogenetic tree of SINE1, SINE2 homologues proteins.
Legend as in Figure 2.

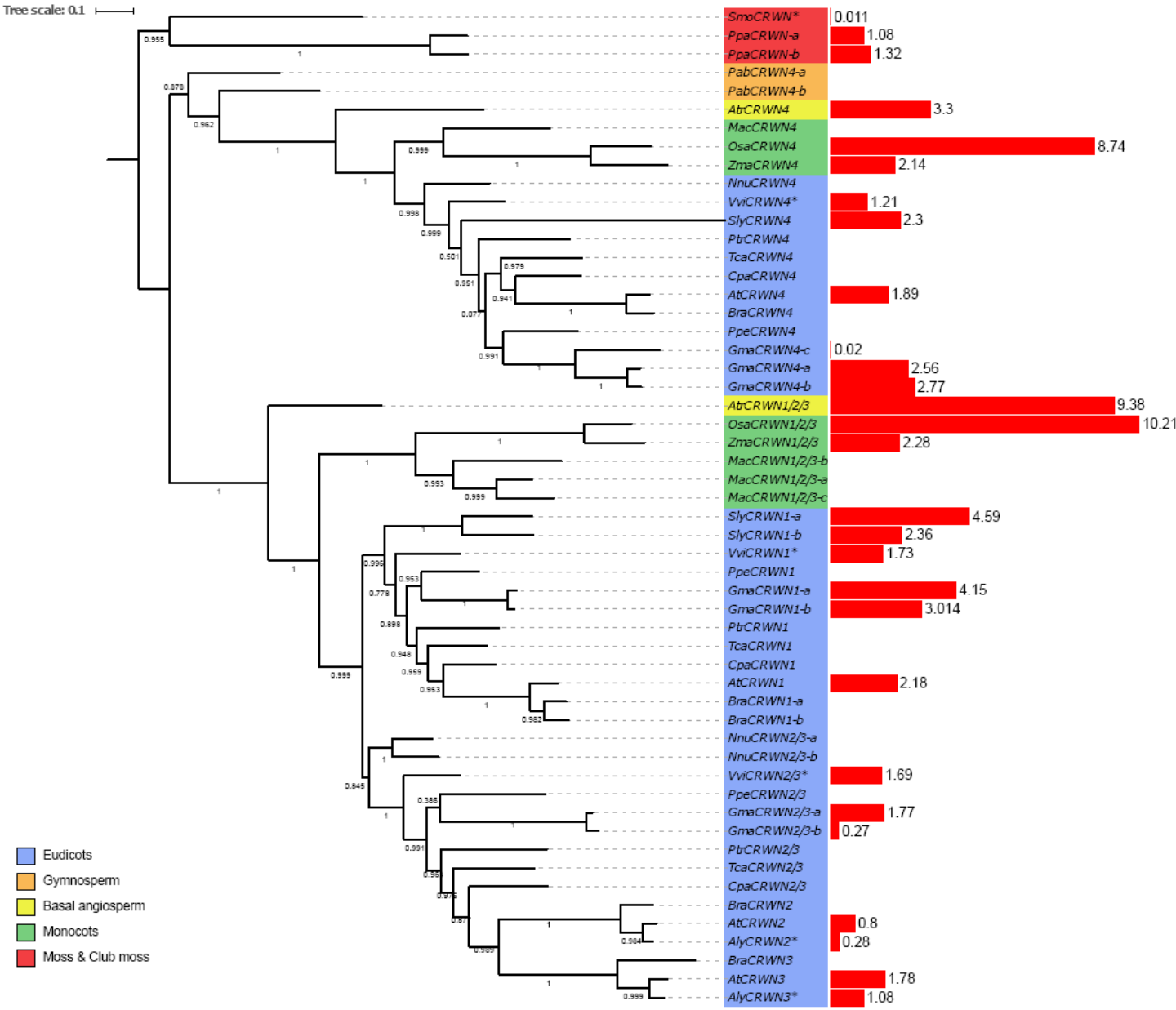


Figure 6: **Phylogenetic tree of CRWN proteins**.
Legend as in Figure 2.

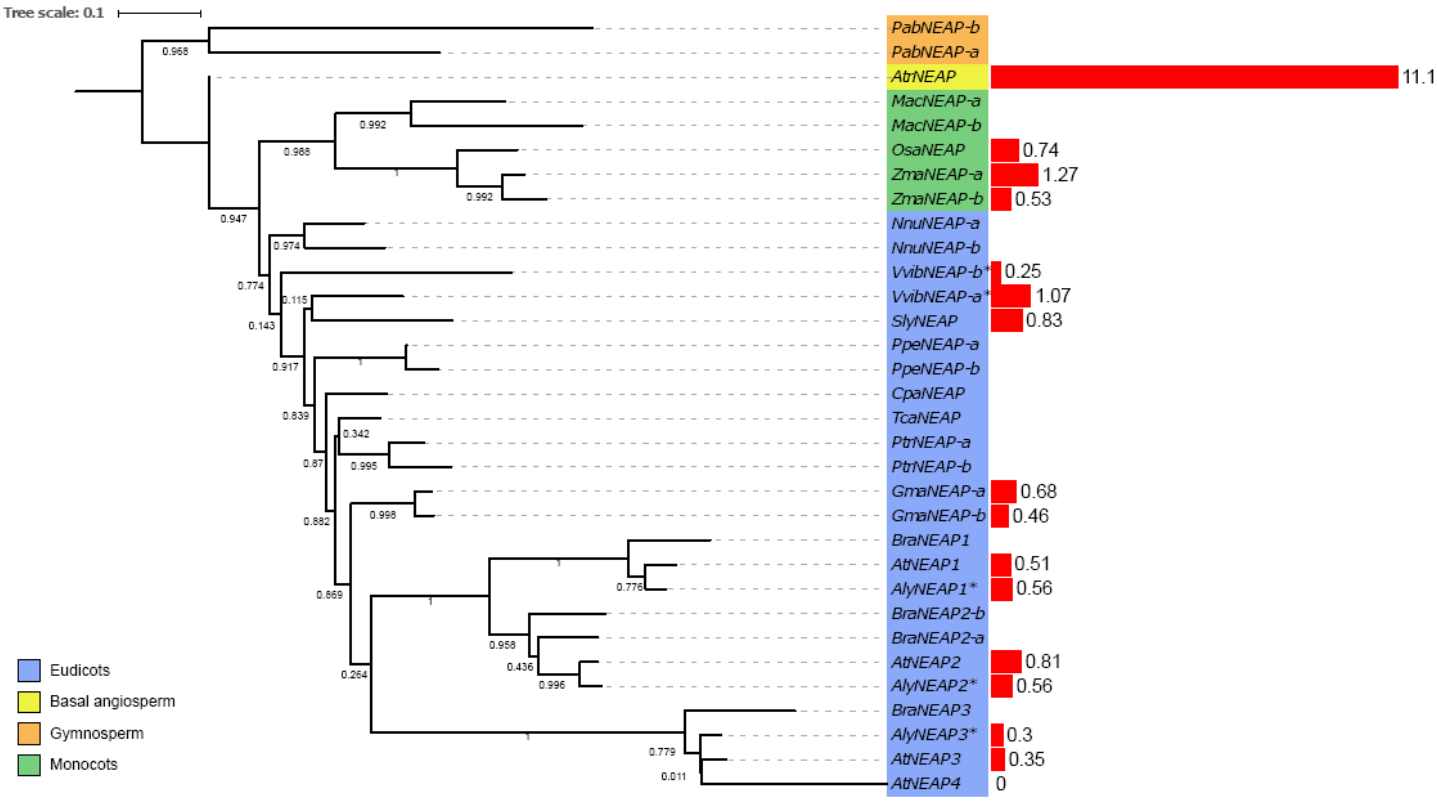


Figure 7: Phylogenetic tree of NEAP proteins.
Legend as in Figure 2.

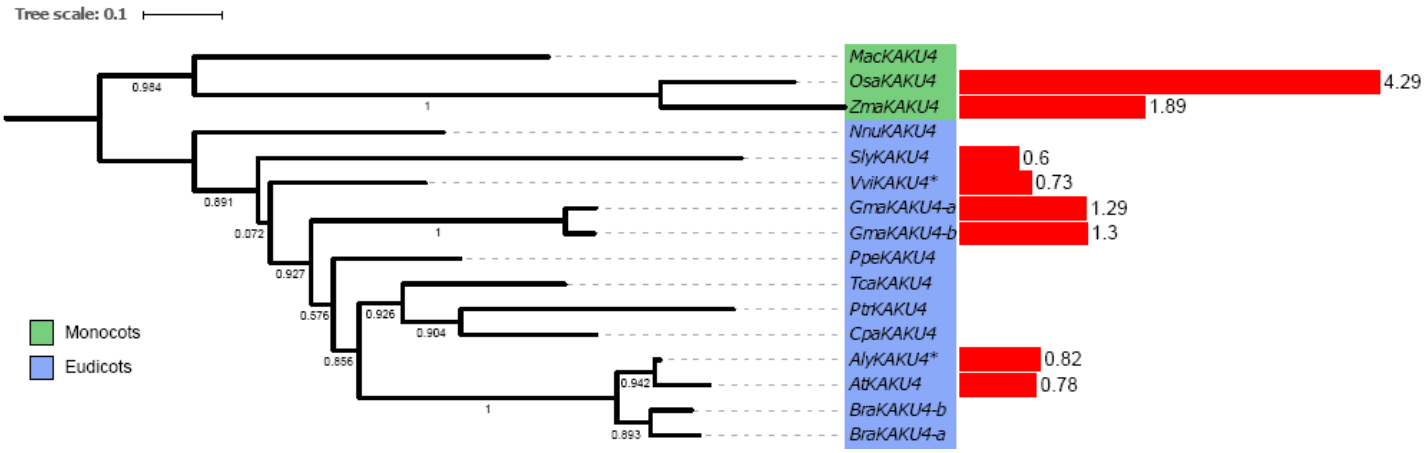


Figure 8: Phylogenetic tree of KAKU4 proteins.
Legend as in Figure 2.