New Phytologist

Research

Evolutionary history of HOMEODOMAIN LEUCINE ZIPPER transcription factors during plant transition to land

Facundo Romani[1], Renata Reinheimer[2], Stevie N. Florent[3], John L. Bowman[3] and Javier E. Moreno[1]

[1]Instituto de Agrobiotecnología del Litoral, Universidad Nacional del Litoral – CONICET, Facultad de Bioquímica y Ciencias Biológicas, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nacional No. 168 km. 0, Paraje El Pozo, Santa Fe 3000, Argentina; [2]Instituto de Agrobiotecnología del Litoral, Universidad Nacional del Litoral – CONICET, Facultad de Ciencias Agrarias, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nacional No. 168 km. 0, Paraje El Pozo, Santa Fe 3000, Argentina; [3]School of Biological Sciences, Monash University, Melbourne, Vic. 3800, Australia

Author for correspondence:
Javier E. Moreno
Tel: +54 342 4511370 ext. 5019
Email: javier.moreno@santafe-conicet.gov.ar

Received: *17 November 2017*
Accepted: *26 February 2018*

New Phytologist (2018)
doi: 10.1111/nph.15133

Key words: development, domain architecture, evolution, HD-Zip, land colonization, transcription factors.

Summary

- Plant transition to land required several regulatory adaptations. The mechanisms behind these changes remain unknown. Since the evolution of transcription factors (TFs) families accompanied this transition, we studied the HOMEODOMAIN LEUCINE ZIPPER (HDZ) TF family known to control key developmental and environmental responses.
- We performed a phylogenetic and bioinformatics analysis of *HDZ* genes using transcriptomic and genomic datasets from a wide range of Viridiplantae species.
- We found evidence for the existence of *HDZ* genes in chlorophytes and early-divergent charophytes identifying several HDZ members belonging to the four known classes (I–IV). Furthermore, we inferred a progressive incorporation of auxiliary motifs. Interestingly, most of the structural features were already present in ancient lineages. Our phylogenetic analysis inferred that the origin of classes I, III, and IV is monophyletic in land plants in respect to charophytes. However, class II *HDZ* genes have two conserved lineages in charophytes and mosses that differ in the CPSCE motif.
- Our results indicate that the HDZ family was already present in green algae. Later, the HDZ family expanded accompanying critical plant traits. Once on land, the HDZ family experienced multiple duplication events that promoted fundamental neo- and subfunctionalizations for terrestrial life.

Introduction

Land plant colonization and radiation are major keystones in the evolutionary history of living organisms, shaping the atmosphere and landscape on Earth to what we know today. This transition was accompanied by morphological, physiological, and genetic changes to cope with the terrestrial environment and its challenging conditions, including: increased CO_2 concentration and light intensity, desiccation, limited nutrient availability, and marked seasonal changes (Kenrick & Crane, 1997; Dahl *et al.*, 2010; Delaux *et al.*, 2012; Delwiche & Cooper, 2015). Biochemical and genomic studies suggest that ancestral charophytes evolved to achieve land colonization (Mikkelsen *et al.*, 2014; Holzinger & Pichrtova, 2016). Charophytes are typically freshwater-living organisms and the closest lineage to land plants. This lineage harbors many innovations essential for aeroterrestrial life and might have lived on land even before Embryophytes (Stebbins & Hill, 1980; Graham *et al.*, 2012; Harholt *et al.*, 2016). At the cellular level, the physiological changes are orchestrated by transcription factors (TFs). It is now clear that plant colonization and radiation was predated by a significant increase in the number of TFs gene families (Holland, 2013; de Mendoza *et al.*, 2013; Catarino *et al.*, 2016). This diversification was followed by another steep rise in the within-class number of TFs in land plant genomes, suggesting a fundamental role of TFs in the adaptation of plants to the new environment, likely through neofunctionalization and subfunctionalization (Zalewski *et al.*, 2013; Hughes *et al.*, 2014; Rensing, 2014; Moghe & Last, 2015).

Until recently, our knowledge on the evolutionary history of TF families during the early radiation of green plants (Viridiplantae) was limited by genomic and transcriptomic resources of model species, such as *Chlamydomonas reinhardtii* and *Physcomitrella patens*. The genome sequences of *Klebsormidium nitens*, an early-diverging charophyte species (Hori *et al.*, 2014), and the liverwort *Marchantia polymorpha*, a basal land plant (Bowman *et al.*, 2017), represented significant advances in the field. Unlike other land plant genomes, the *M. polymorpha* genome showed low redundancy of regulatory genes, such as TFs, in contrast to other gene families related to structural and metabolic traits (Bowman *et al.*, 2017). A recent study using fully sequenced genomes showed that 39 of the 48 plant TF families were already encoded in the genome of *K. nitens*, and the vast

majority of them were incorporated before land colonization during the Precambrian Eon (Catarino *et al.*, 2016; Harholt *et al.*, 2016). At this time, there are several transcriptomic and genomic projects aiming to fill the information gap for chlorophytes and charophytes species, including 'The green algal tree of life' and '1kp' projects (Matasci *et al.*, 2014; Cooper & Delwiche, 2016). These projects might help to elucidate the closest living chlorophyte class to charophytes, since it is now clear that Mesostigmatophyceae and Chlorokybophyceae are early-divergent charophyte classes (Lemieux *et al.*, 2007).

The homeobox TF superfamily is found in all eukaryotic organisms, and is characterized by the presence of a homeodomain (HD), a conserved stretch of 60 amino acid residues that fold into a three-helix DNA-interacting structure (Burglin & Affolter, 2016). In land plants, the HD superfamily is classified into 11 families according to conserved domains: KNOX, BEL, LD, PINTOX, HDZ, WOX, PLINC, NDX, SAWADEE, PHD, and DDT (Mukherjee *et al.*, 2009). The physiological role of these proteins is diverse, spanning from developmental roles to environmental stress responses, exemplified by the WOX, KNOX/BELL, and HDZ families all having a major impact on *Arabidopsis* development (Capella *et al.*, 2015).

In this work, we centered our interest on the evolution of the homeodomain-leucine zipper (HDZ) family. There are four classes of HDZ (I–IV), each playing specific roles in plant development and physiology (Ariel *et al.*, 2007). All classes are characterized by a DNA-binding HD followed by a leucine zipper (LZ) domain required for intra-class dimerization and DNA binding (Ariel *et al.*, 2007), though inter-class dimers may also be plausible (Brandt *et al.*, 2014). The LZ domain is a regular arrangement of aliphatic amino acid residues (such as leucine, methionine, valine, and isoleucine) in the fourth position of heptad repeats (Deppmann *et al.*, 2004). Interestingly, each HDZ class shows a specific number of heptad repeats: whereas class I HDZ (C1HDZ) sequences present between six and five repeats, class II HDZ (C2HDZ) proteins have four, and most of class III HDZ (C3HDZ) and class IV HDZ (C4HDZ) proteins show six (Brandt *et al.*, 2014). Additional class-specific auxiliary motifs with lower conservation levels than HD and LZ can also be found, for example; the transactivation activity of C1HDZ relies on a conserved aromatic, large hydrophobic, acidic context (AHA)-like motif located downstream of the LZ domain (Capella *et al.*, 2014). Furthermore, C2HDZ proteins are characterized by the presence of two exclusive motifs: the C-terminal CPSCE sequence, and the N-terminus ZIBEL-like motif, a short tag of *c.* 10 amino acid residues (Mukherjee *et al.*, 2009). Finally, C3HDZ and C4HDZ proteins show a conserved START/SAD domain required for TF activity (Ponting & Aravind, 1999; Schrick *et al.*, 2004), and the C3HDZ also bears a unique MEKHLA domain that resembles the fungal PAS domain (Mukherjee & Burglin, 2006). These auxiliary motifs can be useful to track the evolutionary history of the family, including gene duplication and gene loss events. Despite their importance, there is still a lack of information regarding their role in the evolution of the HDZ family and their association to the functional role of each class.

The HDZ proteins bind to a pseudopalindromic sequence whose core nucleotides are composed of AATNATT (Palena *et al.*, 2001). These results were recently confirmed using high-throughput approaches (Franco-Zorrilla *et al.*, 2014; O'Malley *et al.*, 2016). However, HDZ proteins of different classes play separate physiological roles in angiosperms, since changes in the native expression levels through ectopic expression, *knock-out*, and *knock-down* plants, induced different phenotypes in *Arabidopsis*. Whereas C1HDZ and C2HDZ were mainly involved in responses to biotic and abiotic stress (Ciarbelli *et al.*, 2008; Romani *et al.*, 2016; Moreno-Piovano *et al.*, 2017), C3HDZ and C4HDZ were characterized in mutants with developmental problems related to shoot and root patterning, including, but not limited to, leaf shape and vasculature organization (McConnell *et al.*, 2001; Emery *et al.*, 2003; Green *et al.*, 2005; Nakamura *et al.*, 2006; Wu *et al.*, 2011). The function of *C3HDZ* genes shows an additional layer of regulation, since it is finely regulated at the posttranscriptional level by microRNAs – miR165 and miR166 – to modulate polar development of the leaf primordium in *Arabidopsis* (Emery *et al.*, 2003; Mallory *et al.*, 2004).

Although the four HDZ classes play different developmental roles in the context of land plants, many of these processes do not exist in charophyte species, and, though initially described in land plants, it has been shown that some of the four HDZ classes had their origin in streptophytes (Floyd *et al.*, 2006; Zalewski *et al.*, 2013). To better understand the evolution of *HDZ* genes, we performed this study using transcriptomic assemblies from recently released databases spanning a wide range of plant taxa, from chlorophytes to land plant species. Using this approach, we identified HD families in Viridiplantae species with the aim of elucidating the origins of the different HDZ classes. Here, we report different aspects of the evolutionary history of the HDZ family, including gene duplication and gene loss events, and the evolution of auxiliary motifs in each class; in particular, we discuss in detail the evolution of C1HDZ and C2HDZ in Viridiplantae.

Materials and Methods

Homeodomain classification in Viridiplantae

HD protein sequences of algal species were obtained from two publicly available transcriptome databases: 'The green algal tree of life' project (https://figshare.com/articles/Green_algal_transcriptomes_for_phylogenetics_and_comparative_genomics/1604778) (Cooper & Delwiche, 2016) and '1kp' project (http://www.cyverse.org/) (Matasci *et al.*, 2014). Algal species from each database are described in Supporting Information Table S1. HD sequences were obtained in a tBLASTN (*e*-value < 0.01) search using the HD consensus sequence reported in plantTFDB 4.0 for HDZ as query (Camacho *et al.*, 2009; Jin *et al.*, 2017). Sequences were translated and filtered using the HMMER (*e*-value < 0.01) model matrix built from a previously reported HD alignment (Finn *et al.*, 2015; Catarino *et al.*, 2016). Reference HD protein sequences from *Arabidopsis thaliana*,

M. polymorpha, *Selaginella moellendorffii*, *P. patens*, *K. nitens*, *Chlamydomonas reinhardtii*, *Ostreococcus tauri*, and *Volvox carteri* were obtained from PLANT TFDB 4.0 (http://planttfdb.cbi.pku.edu.cn/) or PHYTOZOME (Goodstein *et al.*, 2012). The resulting sequences were aligned using the MAFFT G-INS-1 iterative method (Katoh & Standley, 2013) and manually trimmed in MEGA7 (Kumar *et al.*, 2016) to obtain a *c.* 85 amino acid region, including an expanded HD. The list was manually curated, eliminating redundant sequences from alternative assemblies and sequences with >50% gaps. Paralog genes were manually removed to avoid overrepresentation of land plant sequences.

The maximum likelihood (ML) analysis was carried out with IQTREE using LG+G4 and state frequencies determined from amino acid matrix and other default parameters (Nguyen *et al.*, 2015). Branch support was tested using a Shimodaira–Hasegawa-like approximate likelihood ratio test. The consensus tree topology was visualized with MEGA7. Using the reference sequences, we classified HD families following current nomenclature (Mukherjee *et al.*, 2009; Catarino *et al.*, 2016).

To further test the identity of those chlorophyte sequences previously not classified as members of HD family, we applied two complementary tests. First, we used the protein sequence as a query in a BLASTP search on the NCBI nonredundant (NR) database to test if the identity of the best-hit from land-plant species, using either the full-length protein or the HD, matched with the same family obtained in the phylogenetic classification. Second, we used the full-length protein sequence to search for conserved domains located outside the HD in INTERPROSCAN using default parameters (Jones *et al.*, 2014).

Multiple sequence alignment, secondary structure prediction and phylogenetic analysis of HDZ

The putative full-length HDZ sequences were aligned and manually trimmed to include both HD and LZ domains. Selected taxa were visualized using JALVIEW v.2 (Waterhouse *et al.*, 2009). Coiled-coil prediction in HDZ sequences was performed using default parameters with jPred4 and MARCOIL (Alva *et al.*, 2016). The phylogenetic analysis of putative HDZ was performed with a region of *c.* 100 amino acid residues containing both domains. Sequences lacking an LZ domain were eliminated. A total of 158 sequences were included in the alignment. The Bayesian phylogenetic inference was carried out using MRBAYES 3.2.6 software (Huelsenbeck & Ronquist, 2001) with parameters of gamma among-site rate variation model with four rate categories and LG amino acid priority model; four runs with four chains each of 50 000 000 generations were calculated using CIPRES resources, achieving a convergence diagnostic value <0.01 (Miller *et al.*, 2010). The consensus tree was obtained with all compatible groups and visualized using FIGTREE v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

A similar protocol was used for the phylogenetic analysis of C1HDZ and C2HDZ protein sequences retrieved from the plantTFDB 4.0 and classified into classes using an ML approach with MEGA7. These sequences were from the following species: *Amborella trichopoda*, *A. thaliana*, *Capsicum annuum*, *M. polymorpha*, *Oryza sativa* subsp. *japonica*, *Populus trichocarpa*, *Pseudotsuga menziesii*, *P. patens* and *Sphagnum fallax*. HDZ sequences from *Bryum argenteum* were retrieved from NCBI's EST database and *Ceratopteris richardii* sequences from NCBI's NR database. The C1HDZ and C2HDZ sequences of *Equisetum giganteum* and other Marchantiophytes species were retrieved from a BLAST search against the *Equisetum* sp. and *Marchantia emarginata* transcriptomes (Trinity Assemblies) in-house (Wickett *et al.*, 2014; Vanneste *et al.*, 2015). The putative ortholog sequences were translated as described before to supplement the coverage of bryophytes sequences in the C1HDZ tree. Each alignment was extended to cover 115 positions surrounding the HDZ domain for C1HDZ and 137 for C2HDZ. In the first case a convergence diagnostic value <0.01 was achieved after 6665 000 generations, and in the second case a value <0.005 after 12 005 000 generations.

Identification of auxiliary domains and motifs

The identification of auxiliary motifs and domains located outside the HDZ was performed using full-length sequences. A Pfam search using a threshold value of 1×10^{-4} was applied to all HDZ sequences to identify the START/SAD domain and MEKHLA motif. For the identification of CPSCE, ZIBEL, and AHA motifs, we used the full-length protein alignment of HDZ and identified those manually using reference sequences with a similarity threshold of 75% (Mukherjee *et al.*, 2009; Arce *et al.*, 2011). The origin of these motifs was inferred using parsimony.

Results and Discussion

Evolution of *HD* genes in Viridiplantae

Land plant colonization and radiation started over 470 Ma (Wellman *et al.*, 2003; Rubinstein *et al.*, 2010). In order to perform a comprehensive phylogenetic analysis of HDZ proteins in plants, we first identified all putative HD proteins in transcriptome databases building a hidden Markov model profile using as a reference the conserved HD sequence reported for a plant TF study covering a similar time range (Catarino *et al.*, 2016). We performed our analysis on recently released green algal transcriptomes (Matasci *et al.*, 2014; Cooper & Delwiche, 2016), including 53 different algal species from chlorophytes, charophytes, and other streptophytes (Table S1).

Using this approach, we identified 286 HD-putative proteins from chlorophytes and charophytes transcriptomes. These sequences were subsequently used to identify the different HD families by applying a phylogenetic analysis. TF families were defined using reference genes based on sequence similarity. We identified in the unrooted ML tree the 11 *HD*-gene families reported (Figs 1, S1; Notes S1) (Mukherjee *et al.*, 2009; Catarino *et al.*, 2016). Interestingly, we found that nine of 11 *HD*-families were already present in chlorophyte transcriptomes (Table S1). We did not find members of *LD* and *NDX* gene families gene families in chlorophytes species, suggesting that they were either lost in chlorophytes species or that they appeared later in evolution,

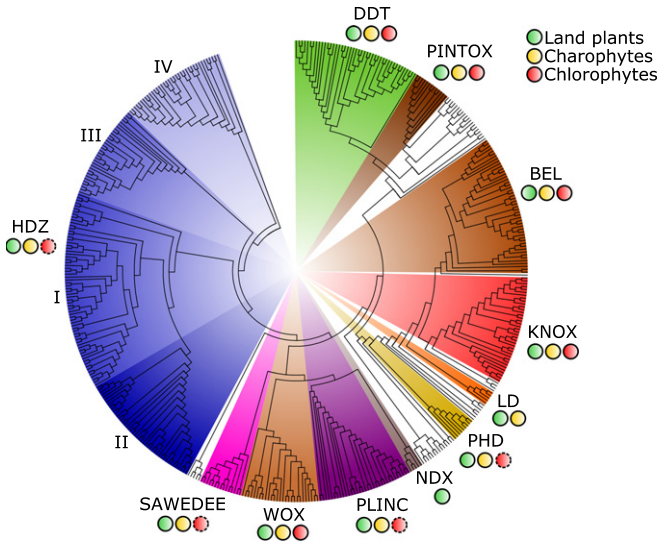

Fig. 1 Phylogeny of *HD* genes in Viridiplantae. The tree was constructed using the amino acid sequence of the HD domain (*c*. 60 aa, Supporting Information Notes S1). Circular cladograms represent the maximum likelihood (ML) analysis of chlorophytes homeodomain proteins. The tree was unrooted ML generated as described in the Materials and Methods section. Subfamilies are highlighted in different colors. The presence of colored circles next to the *HD* gene family indicates the presence of such a protein family in chlorophytes (red), charophytes (yellow), and land plants (green). Circles with dotted lines represent the cases where the origin of the family in chlorophytes was not reported before. The fully annotated tree is presented in Fig. S1.

since these gene families were absent in chlorophytes but present in charophytes in the case of the *LD* family and only present in land plants for the *NDX* family (Fig. 1). More importantly, our analysis supported the presence of the *PHD*, *SAWADEE*, *PLINC* and *HDZ* gene families in chlorophytes (Fig. 1).

Since we are aware of possible contamination of transcriptomic datasets (Laurin-Lemay *et al.*, 2012), we further confirmed the identity of these novel candidates using additional tests. Concisely, we searched for the conservation of the HD and auxiliary domains using the newly identified sequence as a query in a BLASTP search on the NCBI NR database. We used either the HD or the full-length protein to test if the best-hit from land plant sequences was consistent with the family obtained in the phylogenetic classification. In the case of PHD, we obtained an ML bootstrap support value of 81, including one chlorophyte sequence in *O. tauri* (Fig. S1). This protein sequence (fgenesh1_pg.C_Chr_18.0001000131) was defined as unclassified in Catarino *et al.* (2016) and as HB-other in PlantTFdb 4.0. The BLASTP search established the best-hit protein as a PHD homolog. Moreover, the PHD family was previously reported in other chlorophyte species, such as *Micromonas pusilla* and *Ostreococcus lucimarinus* (Hanschen *et al.*, 2016). Regarding the novel SAWADEE family member, we identified one sequence from *Nephroselmis pyriformis* with a bootstrap support value of 98 (Fig. S1). The BLASTP and best-hit supported this phylogenetic classification. Another putative member of the SAWADEE family was previously reported in *O. tauri* (Bowman *et al.*, 2017). We also found one chlorophyte sequence from *Hormotilopsis gelatinosa* belonging to the PLINC family also supported by the BLASTP best-hit analysis. We performed the same screen on the five putative HDZs from chlorophytes: one from *Watanabea reniformis* and three from *Prasiolopsis* sp., both species from the Trebouxiophyceae class, and finally one from *Monomastix opisthostigma* identified in the phylogeny with a support value of 71. In the case of *M. opisthostigma* and *W. reniformis*, the best hit was in accordance with the classification. This was not the case for the sequences belonging to *Prasiolopsis* sp., where we suspect a contamination given their high sequence resemblance to fungal proteins. Intriguingly, we failed to find auxiliary domains in the chlorophyte HDZ sequences using INTERPROSCAN. The identified sequences might be true homologs that have not yet incorporated the auxiliary domains that could point to a stepwise acquisition of these domains during the evolutionary process.

Our data mining strategy proved to be efficient for the detection of *HDZ* genes; for example, no false-positive and no false-negative sequences were found in *K. nitens* when tested against the genome. However, we found several inconsistencies in the conservation of gene families among species from the same subphylum (Table S1) and between transcriptomes datasets. The reason for this might be related to the combination of low gene expression that generated fragmentary sequences, and the occurrence of multiple gene loss events in chlorophytes. Taken together, our phylogenetic analysis of the HD-family suggests that a major divergence in TF families took place early in the evolution of Viridiplantae species, before the diversification of charophytes and, therefore, before plant land colonization. In particular, we identified HDZ family members in chlorophyte species (Fig. 1).

Domain structure and conservation in HDZ proteins of streptophytes

The LZ domain of HDZ proteins is characterized by the presence of heptad repeats with a leucine residue in the center position. This attribute appeared to be conserved even in HDZ proteins from early charophytes species, such as *K. nitens* (Fig. 2). We explored the conservation of both HD and LZ domains between the different protein clades. We assembled a dataset of 158 HDZ sequences, including 110 putative HDZ sequences identified in this work and 48 reference HDZ proteins (Notes S2). The protein alignment showed a clear conservation of the canonical –L(6X)L– repeat for the LZ domain of streptophyte sequences (Fig. 2). Even more, each HDZ class showed the expected number of heptad repeats. On the other hand, chlorophytes HDZ showed a lower conservation of the LZ region compared with the conservation of the HD domain. Only the LZ from *M. opisthostigma* showed conserved positions of leucine residues (Fig. 2). Other chlorophyte sequences showed uneven conservation of leucine positions. Moreover, using prediction software tools, the formation of a coiled-coil domain was only predicted for the LZ of *M. opisthostigma*. For this reason, only this sequence was included in later analysis. Our discovery of putative HDZ in chlorophytes showing high HD similarity with low LZ conservation might reflect either the divergence of orphan genes in these

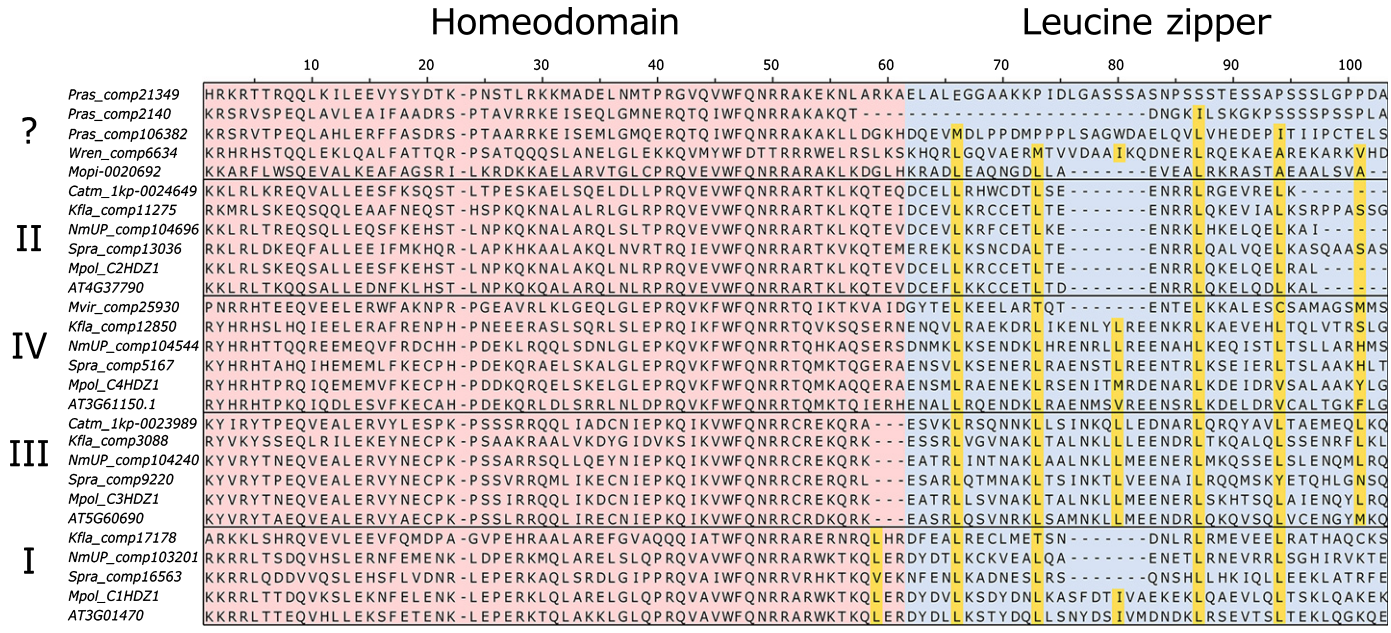

Fig. 2 Sequence alignment of HDZ domains comparing the four classes of HDZ proteins in streptophytes and chlorophytes. The sequence alignment was performed using MAFFT and manually trimmed. Colored leucine residues represent conservation in the alignment. The shaded boxes in pink and blue indicate the HD and LZ domains respectively. Five to seven leucine residues are shown in this alignment. Taxon abbreviations in alphabetical order: Catm, *Chlorokybus atmophyticus*; Kfla, *Klebsormidium nitens*; Mvir, *Mesostigma viridae*; Mopi, *Monomastix opisthostigma*; Mpol, *Marchantia polymorpha*; Pp, *Physcomitrella patens*; Pras, *Prasiolopsis* sp.; NmUP, *Nitella mirabilis*; Spra, *Spirogyra pratensis*; Wren, *Watanabea reniformis*. Each class also includes an *Arabidopsis thaliana* gene.

species, or it could also be evidence of a stepwise evolution of the LZ motif in the gene family.

The four classes of HDZ transcription factors are found in early streptophytes

Since large phylogenetic trees with minimal homologous characters increase the likelihood that homoplasy could confound the relationships, we performed a new analysis using only high-confident HDZ sequences to understand the evolutionary relationships between HDZ proteins in streptophytes. We used the complete HDZ sequences to construct a new phylogenetic tree using a Bayesian approach. We rooted the tree to the only chlorophyte sequence identified as a full HDZ protein. Using this approach, we identified the four HDZ (I–IV) classes based on the topology of the tree, branch lengths, bootstrap values, and visual scrutiny of the primary sequences (Fig. 3; Notes S2). We used the intron–exon structure of fully sequenced genomes as a complementary criterion that may support the classification using the phylogenetic analysis (Fig. S2; Methods S1). Additionally, we rooted the tree to a PHD and also to a C3HDZ protein in order to confirm the classification of the root part of the tree; for example, the chlorophyte HDZ sequence. The resulting trees supported similar results to Fig. 3.

Based on the Bayesian tree, we inferred that *C1HDZ*, *C3HDZ*, and *C4HDZ* land plant genes are monophyletic groups with high support values for *C3HDZ* and *C4HDZ* (> 79%, Fig. 3). Only one member of each class appeared to be inherited by land plants from charophycean algae, with the exception of *C2HDZ* (Fig. 3). Furthermore, we inferred that *C3HDZ* and *C4HDZ* have a common ancestor with high support value (100%) and each class is highly supported (100%), and monophyly has been previously reported for these classes (Floyd *et al.*, 2006; Zalewski *et al.*, 2013). These results are also consistent with a previous report based on genomic sequences from Viridiplantae and red algal species (Mukherjee *et al.*, 2009), and are in agreement with the architecture of the HD domain of *C3HDZ* genes that appeared to be derived relative to the other three classes (Floyd *et al.*, 2006; Zalewski *et al.*, 2013). In our study, both *C3HDZ* and *C4HDZ* were well sustained in the tree (100% and 74% respectively), and the subtrees showed an overall consistency with plant phylogeny, including a basal charophyte at the base of the subtree (Fig. 3). It is interesting to note that the evolution of intron–exon structure for *C3HDZ* and *C4HDZ* genes was significantly different. We observed that *C4HDZ* gene structure is relatively conserved from *K. nitens* to *Arabidopsis* with *c.* 10 exons. On the other side, *C3HDZ* genes have undergone a strong evolution of gene organization, changing from 13 exons in *K. nitens* to a more complex gene structure with *c.* 18 exons in *Marchantia* and *Arabidopsis* (Fig. S2).

While there is no doubt that C1HDZ is monophyletic in land plants, support values were low using these parameters in the phylogeny, and were only acceptable for a small quantity of charophytes sequences closer to the land plant lineage (65%). Charophyte sequences closer to C1HDZ were named class-I-like in the tree (Fig. 3). We found a clade of divergent HDZ sequences close to both C1HDZ and C2HDZ, but difficult to classify specifically into either. Together with the fact that our

New Phytologist

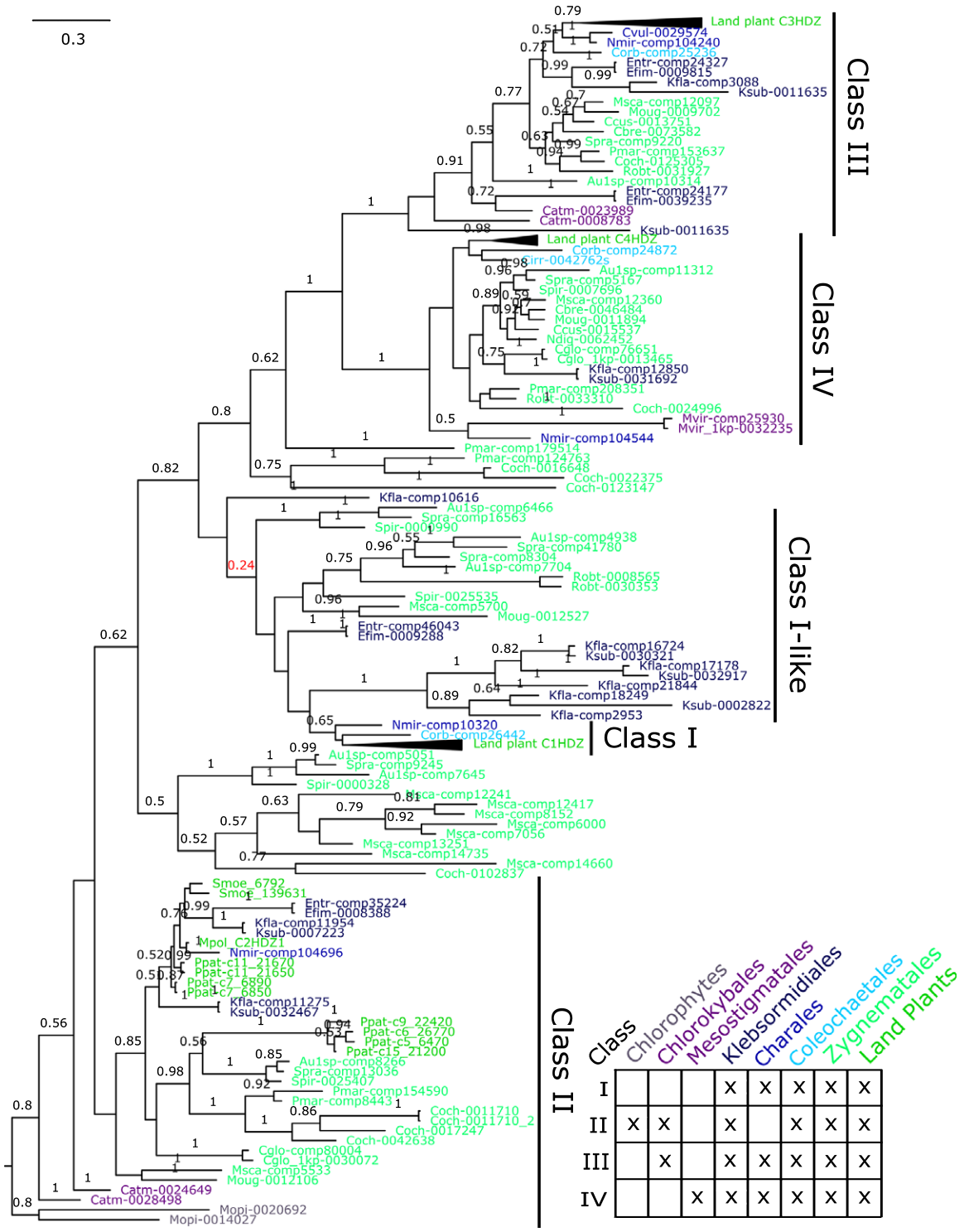

Fig. 3 Phylogeny of HDZ proteins. Bayesian phylogram of streptophytes HDZ protein sequences rooted with *Monomastix opisthostigma* HDZ. Bayesian probability values are shown; values < 0.50 were omitted. Scale bar indicates number of changes per site. Tree was constructed using amino acid sequences of the HDZ domain (Supporting Information Notes S2). The four classes of HDZ are indicated at the right of the tree. Monophyletic land plant taxa were collapsed for C1HDZ, C3HDZ, and C4HDZ for display convenience. The tree is accompanied by a table to summarize the presence (X) or absence (empty) of the class in the plant family. Plant divisions are color coded depending on their taxonomy classification. Chlorophytes (grey); charophytes are divided into classes: Mesostigmatales and Chlorokybales (magenta), Klebsormidiales (dark blue), Charales (blue), Coleochaetales (light blue), Zygnematales (green), and finally land plants (dark green). Taxon abbreviation: Au1sp, *Spirogyra* sp.; Catm, *Chlorokybus atmophyticus*; Cbre, *Cylindrocystis brebissonii*; Ccus, *Cylindrocystis cushleckae*; Cglo, *Chaetosphaeridium globosum*; Cirr, *Coleochaete irregularis*; Coch, *Coleochaete orbicularis*; Corb, *Coleochaete orbicularis*; Cvar, *Chlorella variabilis*; Efim, *Entransia fimbriata*; Entr, *Entransia* sp.; Kfla, *Klebsormidium nitens*; Ksub, *Klebsormidium subtile*; Mopi, *M. opisthostigma*; Moug, *Mougeotia* sp.; Msca, *Mougeotia scalaris*; Ndig, *Netrium digitus*; Nmir, *Nitella mirabilis*; Pmar, *Penium margaritaceum*; Robt, *Roya obtusa*; Spir, *Spirogyra* sp.; Spra, *Spirogyra pratensis*.

study included sequences spanning a broad evolutionary time, we believe that these divergent sequences are responsible for the significant drop in the support values of these branches. Among Zygnematales class-I-like sequences, there were some with high homology to the C1HDZ of land plants, whereas others showed substitutions on key amino acid residues. Additionally, we performed subsequent phylogenetic approaches using ML methods, changing the among-site rate variation, the amino acid substitution model, and reducing the taxa members of each class, but did not improve the trees and arrived at similar results (data not shown). We hypothesize that these events took place later in evolution and did not play a role in the evolution of HDZ that contributed to land plant transition. Moreover, the intron–exon structure of the *C1HDZ* genes of Klebsormidiales showed a particular gene architecture, with an intron located within the HDZ domain that was no longer conserved in *Marchantia* or other land plants (Fig. S2).

It is important to note that the C2HDZ subtree was not fully coincident with plant phylogeny. Interestingly, in addition to *M. opisthostigma*, we identified HDZ proteins belonging to the most basal charophycean species: *Mesostigma viridae* and *Chlorokybus atmophyticus* (Fig. 3). Based on phylogenetic tree inferences, these proteins were associated with different HDZ classes. Whereas the *M. opisthostigma* HDZ resembled a C2HDZ, the *M. viridae* sequence was related to a C4HDZ and *C. atmophyticus* to C2HDZ and C3HDZ sequences (Fig. 3). Thus, the most parsimonious explanation would be that a class-II-like HDZ protein initially evolved in chlorophytes and was later inherited by streptophytes. This hypothesis is consistent with the phylogeny of plants, and is also supported by the conservation of a *C2HDZ* in *C. atmophyticus*. During the evolution of streptophytes, other classes diverged, probably first into a *C3HDZ* and *C4HDZ*, who have a common ancestor and are present in the transcriptome of basal charophyte species, and finally the *C1HDZ* during Klebsormidiophyceae class evolution. This resembles a similar pattern of HD evolution proposed by Mukherjee *et al.* (2009).

We also found two different *C2HDZ* sequences in *P. patens*, suggesting at least one gene duplication event occurred early in charophycean algae evolution that was conserved in some species of land plants, including liverwort and mosses.

The phylogenetic inference also suggested that the four classes of HDZ are present in Klebsormidiales (Fig. 3). This result supports the thesis that these classes diverged early in the evolution of charophytes before or during land colonization. In addition, at least one member of each class appeared to be conserved in Charales and Coleochaetales. Taken together, our data suggest that HDZ proteins appeared during early chlorophyte evolution, and diverged into the four classes during early charophyte evolution.

Origin and evolution of auxiliary motifs in HDZ

We searched for the presence and evolution of auxiliary motifs located outside the HDZ domain using the alignment of the complete protein sequences. The origin of these motifs was inferred by parsimony. In the first place, we looked for the START/SAD domain found only in C3HDZ and C4HDZ (Schrick *et al.*, 2004). In our dataset, the START/SAD domain was already present in *Klebsormidium* and well conserved among virtually all the sequences identified as C3HDZ and C4HDZ proteins (Fig. 4). We inferred that the incorporation of this domain to the C-terminal part of C3HDZ and C4HDZ proteins occurred early in the evolution of streptophytes and before the divergence of Klebsormidiales. Likewise, the MEKHLA domain, exclusive to C3HDZ, was found only in C3HDZ of *Klebsormidium* species and conserved later in evolution (Fig. 4).

Instead, members of the C2HDZ are characterized by the presence of the CPSCE sequence at the C-terminal domain (Ariel *et al.*, 2007). In this regard, our phylogenetic analysis revealed the existence of two clades of C2HDZ that differ by the presence or absence of a CPSCE motif (see later, Fig. 7). These two lineages have a common origin, with duplication before the divergence of Klebsormidiales. The CPSCE-less lineage is only conserved in charophytes and mosses before a subsequent loss in vascular plants (see later, Fig. 7). The ZIBEL-like motif is a sequence of *c.* 10 amino acid residues identified in the N-terminus of C2HDZ proteins (Mukherjee *et al.*, 2009). Unlike the CPSCE sequence, the ZIBEL-like motif was found in Klebsormidiales and conserved in all C2HDZs (Fig. 4).

It was shown that the transactivation activity of C1HDZ relies on a conserved AHA-like motif (Capella *et al.*, 2014). According to our phylogenetic inference, this motif has an origin before the divergence of Charales and is well conserved in other charophytes (Fig. 4). In land plants, this AHA motif is particularly conserved in bryophytes but is later degenerated in some angiosperm lineages.

In addition to these structural features, the physiological role of C3HDZ TFs is finely tuned by microRNA action (Floyd *et al.*, 2006). *Arabidopsis C3HDZ* genes contain the target sequence for miRNA165/166 in the SAD domain, and it has already been shown that *P. patens* and *M. polymorpha C3HDZ* genes are regulated by miR166 (Floyd & Bowman, 2004; Yip *et al.*, 2016). In this context, we wondered what the conservation

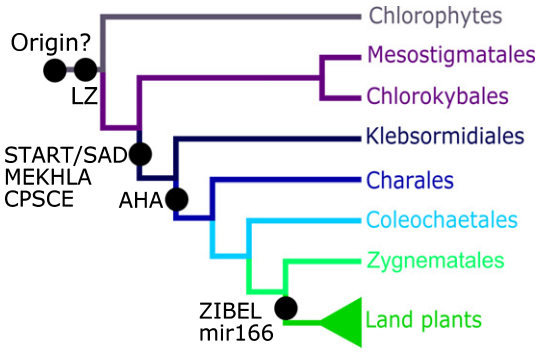

Fig. 4 Hypothetical evolutionary model for auxiliary motifs located outside the HDZ domain in the gene family. Black circles represent the event of incorporation of each motif. Major taxonomic groups are shown at the tip branch of the tree. Branches are colored depending on their taxonomy classification: chlorophytes; charophytes are divided into classes: Mesostigmatales and Chlorokybales, Klebsormidiales, Charales, Coleochaetales, Zygnematales, and finally land plants.

New Phytologist

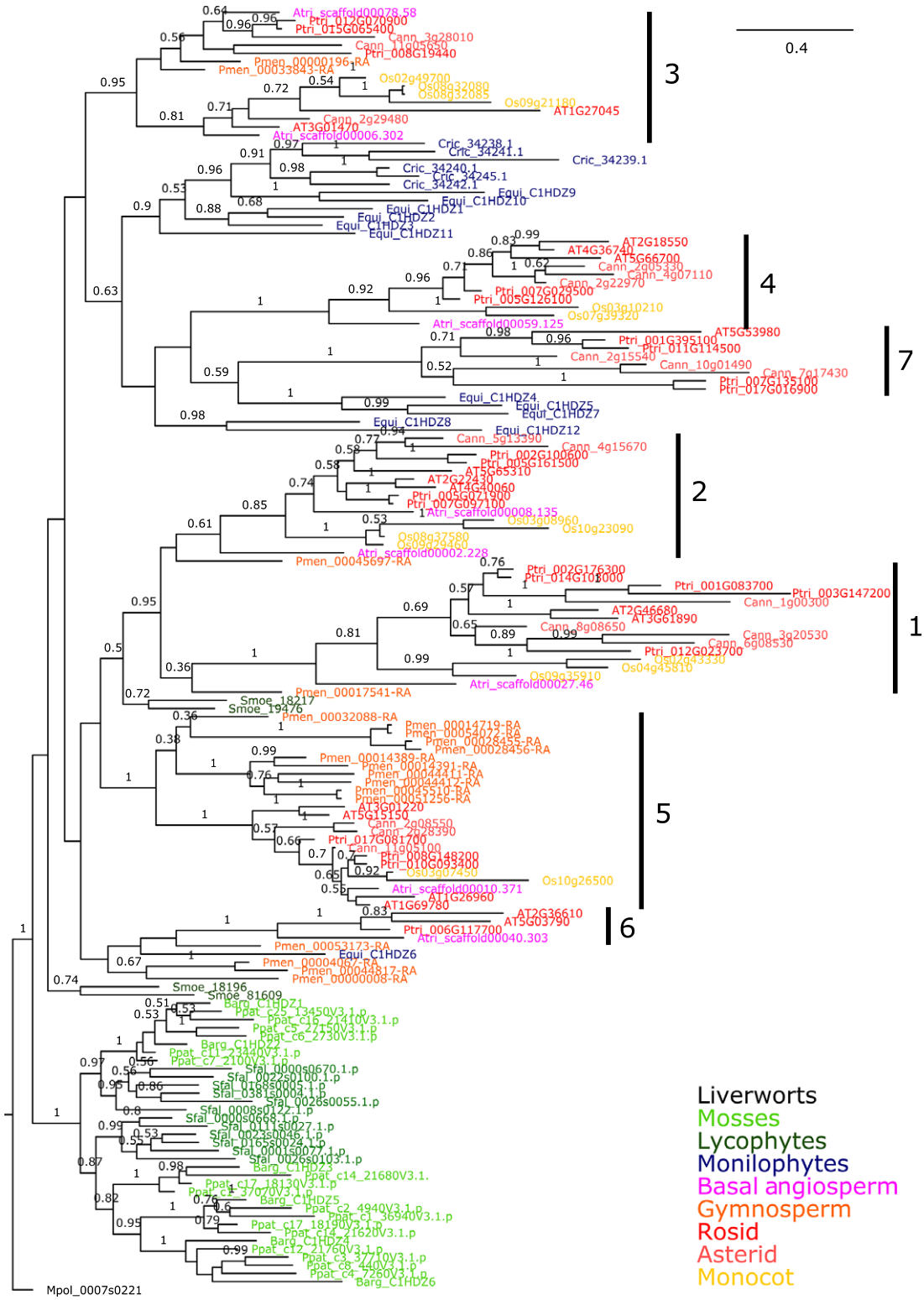

Fig. 5 Bayesian phylogram of *C1HDZ* genes from land plants. Tree was constructed using amino acid alignment (Supporting Information Notes S3). Numbers at branches indicate posterior probability values. Scale bar indicates number of changes per site. Seven clades (1–7) are highlighted in the tree. Taxa are color coded according to major plant classification. AT, *Arabidopsis thaliana* (red, rosid); Atri, *Amborella trichopoda* (fuchsia, basal angiosperm); Barg, *Bryum argenteum* (green, bryophyte); Cann, *Capsicum annuum* (pink, asterid); Cric, *Ceratopteris richardii*; Equi, *Equisetum giganteum* (dark blue, monilophytes); Mpol, *Marchantia polymorpha* (black, liverwort); Os, *Oryza sativa* (yellow, monocot); Pmen, *Pseudotsuga menziesii* (orange, gymnosperm); Ppat, *Physcomitrella patens* (green, bryophyte); Ptri, *Populus trichocarpa* (red, rosid); Smoe, *Selaginella moellendorffii* (dark green, lycophytes); Sfal, *Sphagnum fallax* (green, bryophyte).

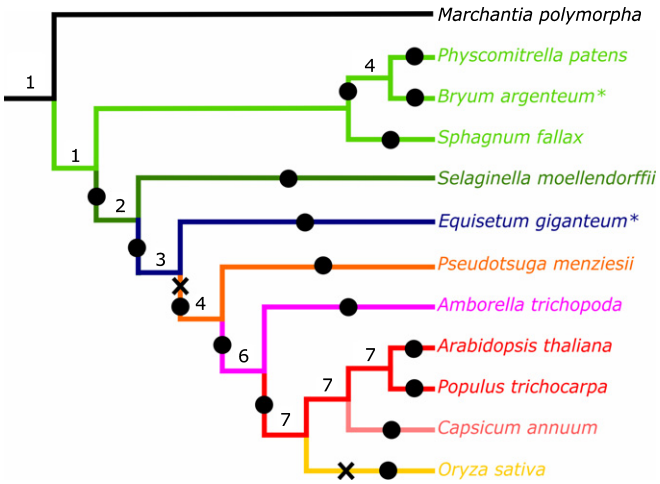

Fig. 6 Hypothetical evolutionary model for *C1HDZ* genes from land plants. Black circles represent gene duplication events, and black crosses represent gene loss events inferred from the phylogenetic analysis. Numbers in the branches indicate the hypothetical number of members of the class in the common ancestor. Representative species of each major taxonomic group are shown at the branch tip. Branches are colored depending on their taxonomy classification: *Marchantia polymorpha* (liverwort); bryophytes including *Physcomitrella patens*, *Bryum argenteum* and *Sphagnum fallax* (bryophytes); *Equisetum giganteum* (monilophytes); *Selaginella moellendorffii* (lycophytes); *Pseudotsuga menziesii* (gymnosperm); *Amborella trichopoda* (basal angiosperm); *Arabidopsis thaliana* and *Populus trichocarpa* (rosid); *Capsicum annuum* (asterid); and *Oryza sativa* (monocot). Asterisks indicate species without a sequenced genome.

level of the miR166 target site was in these genes of charophytes and early land plants. The nucleotide sequence alignment showed a relatively low conservation of the microRNA target sequence in charophytes (Fig. S3; Methods S1). Moreover, in the best case, there were six additional mismatches in the alignment of the microRNA and its target sequence, something already shown for *Chara coralline* (Floyd *et al.*, 2006). This suggests that the miR166 binding sequence, as we know it today, is not conserved in early charophyte species, indicating that this miRNA target site evolved in the ancestral land plant.

Evolution of *C1HDZ* genes in land plants

The evolution of *C3HDZ* and *C4HDZ* genes was well described before, where the expansion of the gene family accompanied vascular plant radiation (Floyd *et al.*, 2006; Zalewski *et al.*, 2013). For this reason, we decided to study in detail the events of gene duplications and gene losses of *C1HDZ* in plants. Since C1HDZ proteins in land plants are part of a monophyletic group (Fig. 3) (Henriksson *et al.*, 2005), and considering that the initial runs to convergence including charophycean sequences were not consistent with plant phylogeny, we performed the analysis using only land plant sequences (Notes S3). In order to understand their evolution in land plants, we analyzed sequences from representative taxa spanning a wide range of model species with fully sequenced genomes along with two transcriptomic sets of bryophyte species to improve their representation in the phylogeny. When the tree was rooted using the liverwort sequence Mp*C1HDZ*, the results were largely consistent with the phylogeny of land plants (Fig. 5). This analysis supported the existence of seven clades. In comparison with previous reports (Henriksson *et al.*, 2005; Arce *et al.*, 2011), clade 1 is equivalent to clade γ, clade 2 to β, clade 4 to δ, clade 5 to α, clade 6 to ε, clade 7 to φ, and clade 3 includes members of clades β and φ according to Henriksson *et al.* (2005). Compared with a more recent study (Arce *et al.*, 2011), it shows the same six clades identified as clades 1–6 and shows an extra clade.

The Bayesian phylogenetic tree indicated that there was a common ancestor to bryophytes and vascular plants that was inherited from a single *C1HDZ* gene (Fig. 6). The evolutionary history of gene duplication events in mosses is independent from those in seed plants. Phylogenetic inference suggested the occurrence of a gene duplication event in lycophytes that was conserved in seed plants (Fig. 6). In addition, we found at least two duplication events in *E. giganteum* (Fig. 6). This horsetail also showed an exclusive clade for this gene class. We also inferred two conserved duplication events before the diversification of gymnosperms, since four classes were present in the transcriptome of the early gymnosperm *P. menziesii* and two additional duplication events were inferred before or early in the evolution of flowering plants since they were already present in the genome of *A. trichopoda* (Fig. 6), considered a basal flowering plant.

The analysis of C1HDZ sequences in monocot species showed interesting singularities (Fig. 6). Clades 6 and 7 were lost in all monocots (Fig. S4), while a closer look into secondarily aquatic monocot species, such as *Lemna minor*, *Spirodela polyrhiza*, and the recently sequenced seagrass *Zostera marina*, suggested that they have also lost gene members of clade 1 (Fig. S4; Notes S4; Methods S1). Additional BLAST searches were thus performed to detect the presence of *C1HDZ* genes in the genomes of other aquatic monocot species available in NCBI databases. We were able to detect one gene of this class in the genome of the aquatic orchid *Dendrobium nobile*, suggesting that the loss was during the re-adaptation of monocots to water (data not shown). *L. minor* appeared to have also lost gene members of clade 4. Based on these phylogenetic inferences, we conclude that *C1HDZ* genes showed a complex evolutionary history with several lineage-specific duplication and loss events.

Evolution of *C2HDZ* genes in streptophytes

Unlike other HDZ classes, *C2HDZ* genes are not monophyletic, indicating that more than one member was inherited from charophytes to land plants. To better understand this, we performed a phylogenetic analysis with a Bayesian approach of all *C2HDZ* genes, including sequences from green algae to land plants species. As shown in Fig. 3, the phylogenetic inference indicated that there are at least two lineages of *C2HDZ* genes. For this reason, we decided to include additional sequences from charophyte species and to root the tree using *Kfla_comp11954*, given that generated fewer inconsistencies than other basal charophytes sequences. Two divergent sequences from Coleochaetales and Charales were eliminated (Notes S5). Although we were able to run this analysis to convergence (split frequency < 0.005), we

New Phytologist

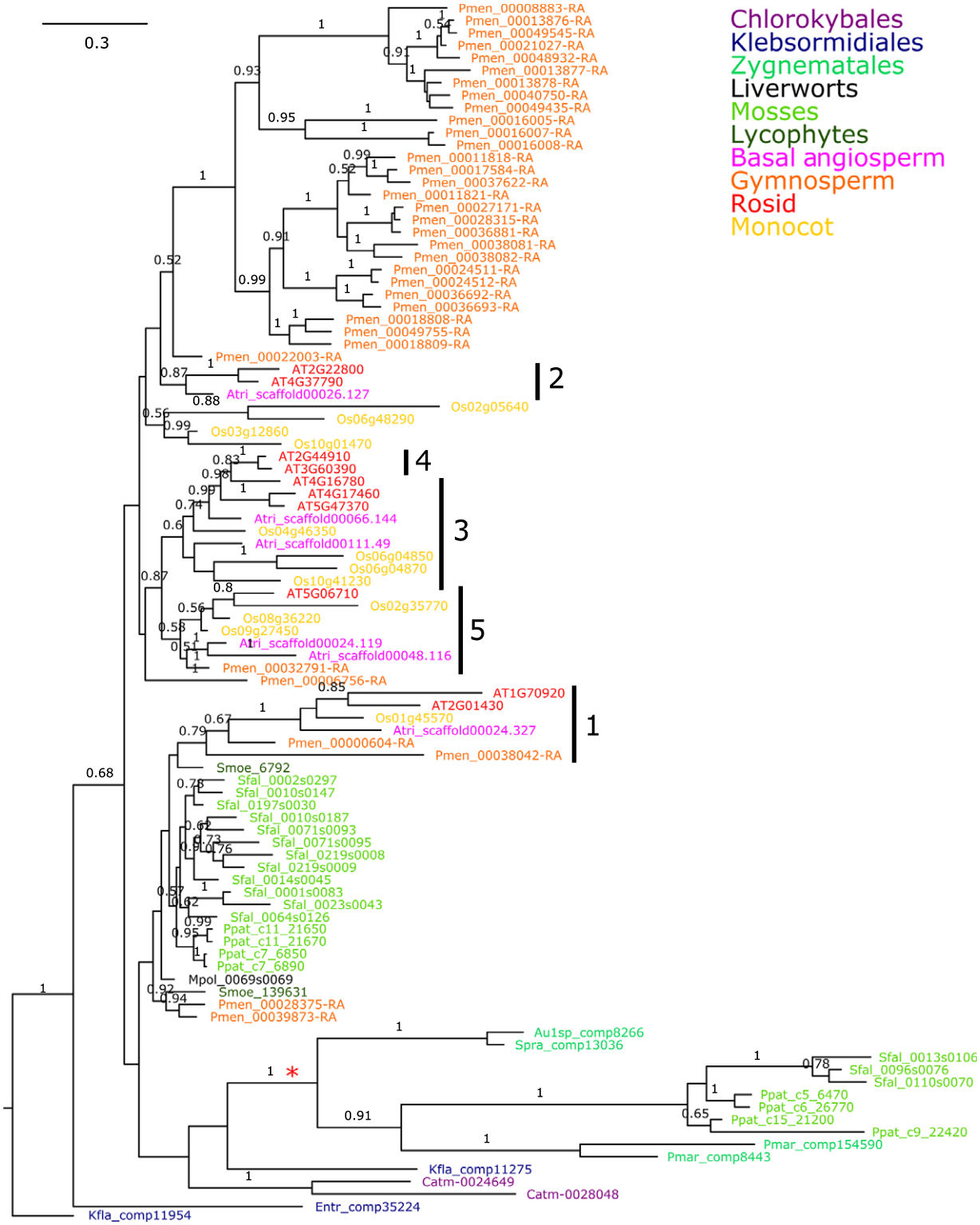

Fig. 7 Bayesian phylogram of *C2HDZ* genes from Viridiplantae species. The tree was constructed using residues of relevant positions as described in Materials and Methods section and is included in supplementary data (Supporting Information Notes S5). Numbers at branches indicate posterior probability values. The red asterisk in the branch denotes the position of the conserved hypothetical loss event of the CPSCE motif. The presence of the CPSCE motif was determined as the loss of conservation of two or more positions in the amino acid sequence. Scale bar indicates number of changes per site. Taxa are color coded according to major plant classification: Chlorokybales (magenta), Klebsormidiales (dark blue), Zygnematales (green), bryophytes (light green), lycophytes (dark green), basal angiosperm (fuchsia), gymnosperm (orange), rosid (red), and monocot plants (yellow). Taxa abbreviations in alphabetical order: AT: *Arabidopsis thaliana*; Atri, *Amborella trichopoda*; Au1sp, *Spirogyra* sp.; Catm, *Chlorokybus atmophyticus*; Entr, *Entransia* sp.; Kfla, *Klebsormidium nitens*; Mpol, *Marchantia polymorpha*; Msca, *Mougeotia scalaris*; Os, *Oryza sativa*; Pmar, *Penium margaritaceum*; Pmen, *Pseudotsuga menziesii*; Ppat, *Physcomitrella patens*; Sfal, *Sphagnum fallax*; Smoe, *Selaginella moellendorffii*; Spra, *Spirogyra pratensis*.

obtained some inconsistencies with plant phylogeny and low supported branches. We obtained a robust tree without a requirement to trim down the alignment or to eliminate more sequences from the analysis (Fig. 7).

Our phylogenetic analysis supported the existence of two different lineages of *C2HDZ* genes that differ in the presence or absence of the CPSCE motif. These two sub-clades were also present in the genome of *K. nitens*. The CPSCE-less lineage was conserved in Zygnematales, and even in mosses. Interestingly, the CPSCE-less lineage appeared to be lost in Marchantiales and the seed plant lineage (Fig. 7). This was further confirmed in transcriptomes from other liverworts species, including *Ricciocarpus natans*, *M. emarginata*, *Sphaerocarpos texanus*, and *Metzgeria crassipilis*, and from the hornwort *Nothoceros* sp. (data not shown; transcriptomic data from Wickett *et al.*, 2014). The intron–exon structure of *C2HDZ* genes also supports the same evolutionary history, since the CPSCE-less lineage showed a different distribution of introns compared with the lineage containing the CPSCE motif (Fig. S2). Similar to previous reports for *C4HDZ* genes (Zalewski *et al.*, 2013), we found a complete absence of introns in all moss *HDZ* genes. Zalewski *et al.* (2013) proposed a mechanism based on reverse transcription to explain this lack of introns in the moss *P. patens*.

Several gene duplication events appeared moss specific, while a single duplication in the lycophyte *S. moellendorffii* appeared conserved in both monilophytes and seed plants (Fig. 8).

In seed plants, the same Bayesian approach inferred an additional conserved duplication event before gymnosperm divergence, and one more before basal angiosperms, giving origin to four conserved clades in angiosperms. To verify these clades, we made a new phylogenetic analysis excluding CPSCE-less sequences (Fig. S5), and including more angiosperm species and monilophytes sequences. Both trees supported similar results, since we did not find any relevant duplication event in monilophytes species (Fig. S6). We also compared our results with previous reports (Ciarbelli *et al.*, 2008; Harris *et al.*, 2011). The clades 1–4 correspond to clades α–δ in Ciarbelli *et al.* (2008). Clade 5 includes protein homologs of HAT14, which were sorted as unclassified in previous studies. Here, clade 5 appeared as a monophyletic group conserved in rosids and asterids. Although it is not fully clear whether or not clades 3 and 4 are products of the same duplication in basal angiosperms, as stated before by Ciarbelli *et al.* (2008) (Fig. S6). On the other hand, most of the gymnosperm sequences were unclassified (Fig. 7, upper part). These sequences might constitute a new clade not present in angiosperms or a divergent group of those clades.

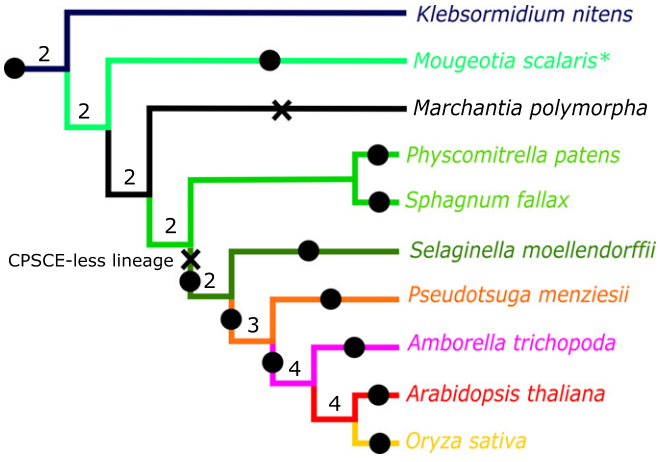

Fig. 8 Hypothetical evolutionary model for *C2HDZ* genes from streptophytes. Black circles represent gene duplication events, and black crosses represent gene loss events inferred from the phylogenetic tree in Fig. 7. Numbers in the branches indicate the hypothetical number of members of the class in the common ancestor. The loss of the CPSCE-less lineage is indicated next to the loss event. Representative species of each major taxonomic group are shown at the branch tip. Branches are colored depending on their taxonomy classification: *Klebsormidium nitens* (Klebsormidiales); *Mougeotia scalaris* (Zygnematales); *Marchantia polymorpha* (liverwort); bryophytes including *Physcomitrella patens* and *Sphagnum fallax*; *Selaginella moellendorffii* (lycophytes); *Pseudotsuga menziesii* (gymnosperm); *Amborella trichopoda* (basal angiosperm); *Arabidopsis thaliana* (rosid); *Capsicum annuum* (asterid); and *Oryza sativa* (monocot). Asterisks indicate species without a sequenced genome.

Conclusion

The recent application of genomic and transcriptomic technologies to study basal lineages of plants is generating powerful data to understand plant evolution and, more specifically, the evolutionary history of gene families. In order to complement genome datasets in a cost-effective way, there are multiple transcriptomic projects spanning a wider evolutionary time (Matasci *et al.*, 2014; Cooper & Delwiche, 2016). Here, we used transcriptomic databases for the identification of *HDZ* genes in chlorophyte and charophyte species. We first analyzed the phylogenetic relationships between members of the whole HD family. We found that nine out of 11 HD families were already present in the genome of chlorophytes species, four more than in previous studies. Our approach based on the use of transcriptomic data significantly expanded the presence of HD TF families in chlorophyte species, including PHD, SAWADEE, PLINC and HDZ families. Although we infer a high rate of gene loss in chlorophytes, it is possible that gene loss in chlorophytes and gene retention in streptophytes played a role in both the divergence of lineages and the colonization of the terrestrial environment respectively. It is still a matter of discussion whether the origin of the family is given by the appearance of a new lineage within a superfamily or by the fusion of protein domains (Long *et al.*, 2003). However, at least in the case of HDZ proteins, the acquisition of auxiliary motifs appeared to have occurred in a stepwise manner (Fig. 4). Our results support the fact that the origin of the majority of HD families took place in Viridiplantae species, and more specifically in chlorophytes. Even more, these results reinforce the idea that most of the diversification of TF families occurred before the Precambrian Eon, and so predated land plant colonization.

Interestingly, the HD of the HDZ protein found in *M. opisthostigma* resembled the HD of C2HDZ proteins. However, we did not find any auxiliary motifs located outside of both HD and LZ domains. Most of these auxiliary motifs appeared before the divergence of Klebsormidiales, with the exception of the ZIBEL and AHA domains in C2HDZ and C1HDZ

respectively. Despite the controversy around basal position of charophytes, *HDZ* sequences were clearly identified in the genome of both basal charophyte species examined, *C. atmophyticus* and *M. viridae*. The most parsimonious interpretation of these results is that *C2HDZ* genes have their origin in chlorophytes. Moreover, the diversity and lack of monophyly in C2HDZ is also consistent with a basal position. This C2HDZ was inherited by streptophytes to diverge later in other classes, first into a C3HDZ and C4HDZ, and finally into the C1HDZ.

This study also revealed interesting gene loss and duplication events that had hitherto not been described. The evolutionary history of C1HDZ in land plants includes four major duplication events that occurred before gymnosperm-angiosperm divergence. C1HDZ evolution in monocots is marked by gene loss events and few duplication events. So far, clades 6 and 7 have been lost in the monocot genomes examined, and in the particular case of secondarily aquatic monocots there is additional loss of gene members of clade 1. It remains unclear what the functional contribution of these clades to the evolutionary process of monocot species is. Considering that several gene members of the clade are involved in abiotic stress responses (Re *et al.*, 2014; Romani *et al.*, 2016), it is speculative to think that these genes are no longer required in an aquatic environment.

Based on our phylogenetic study, the C2HDZ lineage did not show a monophyletic origin in land plants. Our analysis showed that at least two members of C2HDZ were inherited by land plants and conserved in mosses. These two subclades differ in the presence or absence of the CPSCE motif. Our phylogenetic results are consistent to infer that the CPSCE-less clade was lost in vascular plants.

Our results demonstrate that the HDZ family family was already present in green algae. However, its expansion accompanied important developmental processes, including multicellularity, polar growth, and shape. Once on land, the HDZ family family experienced multiple duplication events that, based on their genetic redundancy in *Arabidopsis* and other model plants, likely underwent neo- and subfunctionalization (Zalewski *et al.*, 2013; Breuninger *et al.*, 2016; Vasco *et al.*, 2016). Ongoing sequencing projects of green algal genomes promise to add important information to the evolution of green algae and shed light on the functional evolution and physiological role of these TFs during the evolution of plants.

Acknowledgements

We thank Professor Charles Delwiche and Dr Endymion Cooper, University of Maryland, College Park, for giving us access to raw transcriptomic data of *Prasiolopsis* and generating important number of transcriptomic datasets used in this study. We also thank the 1kp Project. We thank helpful discussions with Dr Agustín Arce related to HMMER analysis. This work was supported by Agencia Nacional de Promoción Científica y Tecnológica (PICT2013-3285 to J.E.M.), Consejo Nacional de Investigaciones Científicas y Técnicas (PIP11220130100267CO to J.E.M.), and Universidad Nacional del Litoral (CAID50220140100011LI to J.E.M.). S.N.F. and J.L.B. were supported by the Australian Research Council (DP170100049). F.R. is a doctoral fellow of CONICET. R.R. and J.E.M. are CONICET career members.

Author contributions

F.R., R.R. and J.E.M. conceived the experiments. F.R., R.R., S.N.F., J.L.B. and J.E.M. performed and analyzed the experiments. F.R., S.N.F., J.L.B. and J.E.M. wrote the paper.

ORCID

Facundo Romani http://orcid.org/0000-0003-3954-6740
John L. Bowman http://orcid.org/0000-0001-7347-3691
Javier E. Moreno http://orcid.org/0000-0001-9763-5325

References

Alva V, Nam SZ, Soding J, Lupas AN. 2016. The MPI Bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Research* **44**(W1): W410–W415.

Arce AL, Raineri J, Capella M, Cabello JV, Chan RL. 2011. Uncharacterized conserved motifs outside the HD-Zip domain in HD-Zip subfamily I transcription factors; a potential source of functional diversity. *BMC Plant Biology* **11**: 42.

Ariel FD, Manavella PA, Dezar CA, Chan RL. 2007. The true story of the HD-Zip family. *Trends in Plant Science* **12**: 419–426.

Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu S, Ishizaki K, Yamaoka S, Nishihama R, Nakamura Y, Berger F *et al.* 2017. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**: 287–304 e215.

Brandt R, Cabedo M, Xie Y, Wenkel S. 2014. Homeodomain leucine-zipper proteins and their role in synchronizing growth and development with the environment. *Journal of Integrative Plant Biology* **56**: 518–526.

Breuninger H, Thamm A, Streubel S, Sakayama H, Nishiyama T, Dolan L. 2016. Diversification of a transcription factor family led to the evolution of antagonistically acting genetic regulators of root hair growth. *Current Biology* **26**: 1622–1628.

Burglin TR, Affolter M. 2016. Homeodomain proteins: an update. *Chromosoma* **125**: 497–521.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Capella M, Re DA, Arce AL, Chan RL. 2014. Plant homeodomain-leucine zipper I transcription factors exhibit different functional AHA motifs that selectively interact with TBP or/and TFIIB. *Plant Cell Reports* **33**: 955–967.

Capella M, Ribone PA, Arce AL, Chan RL. 2015. Homeodomain–leucine zipper transcription factors: structural features of these proteins, unique to plants. In: Gonzalez DH, ed. *Plant transcription factors*. Boston, MA, USA: Academic Press, 113–126.

Catarino B, Hetherington AJ, Emms DM, Kelly S, Dolan L. 2016. The stepwise increase in the number of transcription factor families in the Precambrian predated the diversification of plants on land. *Molecular Biology and Evolution* **33**: 2815–2819.

Ciarbelli AR, Ciolfi A, Salvucci S, Ruzza V, Possenti M, Carabelli M, Fruscalzo A, Sessa G, Morelli G, Ruberti I. 2008. The Arabidopsis homeodomain-leucine zipper II gene family: diversity and redundancy. *Plant Molecular Biology* **68**: 465–478.

Cooper E, Delwiche C. 2016. *Green algal transcriptomes for phylogenetics and comparative genomics.* [WWW document] https://dx.doi.org/10.6084/m9.figshare.1604778 [accessed 5 January 2017].

Dahl TW, Hammarlund EU, Anbar AD, Bond DP, Gill BC, Gordon GW, Knoll AH, Nielsen AT, Schovsbo NH, Canfield DE. 2010. Devonian rise in atmospheric oxygen correlated to the radiations of terrestrial plants and large predatory fish. *Proceedings of the National Academy of Sciences, USA* **107**: 17911–17915.

Delaux P-M, Nanda AK, Mathé C, Sejalon-Delmas N, Dunand C. 2012. Molecular and biochemical aspects of plant terrestrialization. *Perspectives in Plant Ecology, Evolution and Systematics* **14**: 49–59.

Delwiche CF, Cooper ED. 2015. The evolutionary origin of a terrestrial flora. *Current Biology* **25**: R899–R910.

Deppmann CD, Acharya A, Rishi V, Wobbes B, Smeekens S, Taparowsky EJ, Vinson C. 2004. Dimerization specificity of all 67 B-ZIP motifs in *Arabidopsis thaliana*: a comparison to *Homo sapiens* B-ZIP motifs. *Nucleic Acids Research* **32**: 3435–3445.

Emery JF, Floyd SK, Alvarez J, Eshed Y, Hawker NP, Izhaki A, Baum SF, Bowman JL. 2003. Radial patterning of Arabidopsis shoots by *class III HD-ZIP* and *KANADI* genes. *Current Biology* **13**: 1768–1774.

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. 2015. HMMER web server: 2015 update. *Nucleic Acids Research* **43**(W1): W30–W38.

Floyd SK, Bowman JL. 2004. Gene regulation: ancient microRNA target sequences in plants. *Nature* **428**: 485–486.

Floyd SK, Zalewski CS, Bowman JL. 2006. Evolution of class III homeodomain-leucine zipper genes in streptophytes. *Genetics* **173**: 373–388.

Franco-Zorrilla JM, Lopez-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. 2014. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences, USA* **111**: 2367–2372.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N *et al.* 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**: D1178–D1186.

Graham LE, Arancibia-Avila P, Taylor WA, Strother PK, Cook ME. 2012. Aeroterrestrial Coleochaete (Streptophyta, Coleochaetales) models early plant adaptation to land. *American Journal of Botany* **99**: 130–144.

Green KA, Prigge MJ, Katzman RB, Clark SE. 2005. *CORONA*, a member of the class III homeodomain leucine zipper gene family in Arabidopsis, regulates stem cell specification and organogenesis. *Plant Cell* **17**: 691–704.

Hanschen ER, Marriage TN, Ferris PJ, Hamaji T, Toyoda A, Fujiyama A, Neme R, Noguchi H, Minakuchi Y, Suzuki M *et al.* 2016. The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. *Nature Communications* **7**: 11370.

Harholt J, Moestrup O, Ulvskov P. 2016. Why plants were terrestrial from the beginning. *Trends in Plant Science* **21**: 96–101.

Harris JC, Hrmova M, Lopato S, Langridge P. 2011. Modulation of plant growth by HD-Zip class I and II transcription factors in response to environmental stimuli. *New Phytologist* **190**: 823–837.

Henriksson E, Olsson AS, Johannesson H, Johansson H, Hanson J, Engstrom P, Soderman E. 2005. *Homeodomain leucine zipper class I* genes in Arabidopsis. Expression patterns and phylogenetic relationships. *Plant Physiology* **139**: 509–518.

Holland PW. 2013. Evolution of homeobox genes. *Wiley Interdisciplinary Reviews: Developmental Biology* **2**: 31–45.

Holzinger A, Pichrtová M. 2016. Abiotic stress tolerance of charophyte green algae: new challenges for omics techniques. *Frontiers in Plant Science* **7**: 678.

Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, Sato S, Yamada T, Mori H, Tajima N *et al.* 2014. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nature Communications* **5**: 3978.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Hughes TE, Langdale JA, Kelly S. 2014. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Research* **24**: 1348–1355.

Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G. 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research* **45**(D1): D1040–D1045.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G *et al.* 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236–1240.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kenrick P, Crane PR. 1997. *The origin and early diversification of land plants – a cladistic study.* Washington, DC, USA: Smithsonian Institution Press.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33**: 1870–1874.

Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology* **22**: R593–R594.

Lemieux C, Otis C, Turmel M. 2007. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biology* **5**: 2.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics* **4**: 865–875.

Mallory AC, Reinhart BJ, Jones-Rhoades MW, Tang G, Zamore PD, Barton MK, Bartel DP. 2004. MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5′ region. *EMBO Journal* **23**: 3356–3364.

Matasci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M *et al.* 2014. Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**: 17.

McConnell JR, Emery J, Eshed Y, Bao N, Bowman J, Barton MK. 2001. Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. *Nature* **411**: 709–713.

de Mendoza A, Sebe-Pedros A, Sestak MS, Matejcic M, Torruella G, Domazet-Loso T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proceedings of the National Academy of Sciences, USA* **110**: E4858–E4866.

Mikkelsen MD, Harholt J, Ulvskov P, Johansen IE, Fangel JU, Doblin MS, Bacic A, Willats WG. 2014. Evidence for land plant cell wall biosynthetic mechanisms in charophyte green algae. *Annals of Botany* **114**: 1217–1236.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, **14**: 1–8.

Moghe GD, Last RL. 2015. Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiology* **169**: 1512–1523.

Moreno-Piovano GS, Moreno JE, Cabello JV, Arce AL, Otegui ME, Chan RL. 2017. A role for *LAX2* in regulating xylem development and lateral-vein symmetry in the leaf. *Annals of Botany* **120**: 577–590.

Mukherjee K, Brocchieri L, Burglin TR. 2009. A comprehensive classification and evolutionary analysis of plant homeobox genes. *Molecular Biology and Evolution* **26**: 2775–2794.

Mukherjee K, Burglin TR. 2006. MEKHLA, a novel domain with similarity to PAS domains, is fused to plant homeodomain-leucine zipper III proteins. *Plant Physiology* **140**: 1142–1150.

Nakamura M, Katsumata H, Abe M, Yabe N, Komeda Y, Yamamoto KT, Takahashi T. 2006. Characterization of the class IV homeodomain-leucine zipper gene family in Arabidopsis. *Plant Physiology* **141**: 1363–1375.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

O'Malley RC, Huang SS, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. 2016. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* **165**: 1280–1292.

Palena CM, Tron AE, Bertoncini CW, Gonzalez DH, Chan RL. 2001. Positively charged residues at the N-terminal arm of the homeodomain are required for efficient DNA binding by homeodomain-leucine zipper proteins. *Journal of Molecular Biology* **308**: 39–47.

Ponting CP, Aravind L. 1999. START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. *Trends in Biochemical Sciences* **24**: 130–132.

Re DA, Capella M, Bonaventure G, Chan RL. 2014. Arabidopsis *AtHB7* and *AtHB12* evolved divergently to fine tune processes associated with growth and responses to water stress. *BMC Plant Biology* **14**: 150.

Rensing SA. 2014. Gene duplication as a driver of plant morphogenetic evolution. *Current Opinion in Plant Biology* **17**: 43–48.

Romani F, Ribone PA, Capella M, Miguel VN, Chan RL. 2016. A matter of quantity: common features in the drought response of transgenic plants overexpressing HD-Zip I transcription factors. *Plant Science* **251**: 139–154.

Rubinstein CV, Gerrienne P, de la Puente GS, Astini RA, Steemans P. 2010. Early middle Ordovician evidence for land plants in Argentina (eastern Gondwana). *New Phytologist* **188**: 365–369.

Stebbins GL, Hill GTC. 1980. Did multicellular plants invade the land? *The American Naturalist* **115**: 342–353.

Schrick K, Nguyen D, Karlowski WM, Mayer KF. 2004. START lipid/sterol-binding domains are amplified in plants and are predominantly associated with homeodomain transcription factors. *Genome Biology* **5**: R41.

Vanneste K, Sterck L, Myburg AA, Van de Peer Y, Mizrachi E. 2015. Horsetails are ancient polyploids: evidence from *Equisetum giganteum*. *Plant Cell* **27**: 1567–1578.

Vasco A, Smalls TL, Graham SW, Cooper ED, Wong GK, Stevenson DW, Moran RC, Ambrose BA. 2016. Challenging the paradigms of leaf evolution: Class III HD-Zips in ferns and lycophytes. *New Phytologist* **212**: 745–758.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191.

Wellman CH, Osterloff PL, Mohiuddin U. 2003. Fragments of the earliest land plants. *Nature* **425**: 282–285.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA *et al.* 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA* **111**: E4859–E4868.

Wu R, Li S, He S, Wassmann F, Yu C, Qin G, Schreiber L, Qu LJ, Gu H. 2011. CFL1, a WW domain protein, regulates cuticle development by modulating the function of HDG1, a class IV homeodomain transcription factor, in rice and Arabidopsis. *Plant Cell* **23**: 3392–3411.

Yip HK, Floyd SK, Sakakibara K, Bowman JL. 2016. Class III HD-Zip activity coordinates leaf development in *Physcomitrella patens*. *Developmental Biology* **419**: 184–197.

Zalewski CS, Floyd SK, Furumizu C, Sakakibara K, Stevenson DW, Bowman JL. 2013. Evolution of the class IV HD-Zip gene family in Streptophytes. *Molecular Biology and Evolution* **30**: 2347–2365.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 Bayesian phylogram of Viridiplantae HD protein sequences.

Fig. S2 Intron–exon structure of *HDZ* genes.

Fig. S3 Nucleotide sequence alignment of the miR166 target site in representative mRNAs of *C3HDZ*.

Fig. S4 Bayesian phylogram of *C1HDZ* genes from monocot species.

Fig. S5 Bayesian phylogram of *C2HDZ* genes from Viridiplantae species excluding CPSCE-less sequences.

Fig. S6 Bayesian phylogram of *C2HDZ* genes from Viridiplantae species without CPSCE-less sequences but including additional angiosperms taxa and monilophytes.

Table S1 List of plant HD transcription factors used in this study

Methods S1 Additional methodological details related to Supporting Information.

Notes S1 Dataset of HD proteins used to identify HD-protein families.

Notes S2 Dataset of putative HDZ proteins used in Fig. 3.

Notes S3 Dataset of HDZ sequences of land plant species.

Notes S4 Dataset of monocot C1HDZ sequences, including aquatic monocot species.

Notes S5 Dataset of C2HDZ sequences used in this study.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.