Title: A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity

Authors: Camille Sayou‡,1,2,3,4, Marie Monniaux‡,1,2,3,4, Max H. Nanao‡,5,6,*, Edwige Moyroud‡,1,2,3,4,a, Samuel F. Brockington7, Emmanuel Thévenon1,2,3,4, Hicham Chahtane1,2,3,4, Norman Warthmann8, b, Michael Melkonian9, Yong Zhang10, Gane Ka-Shu Wong10, 11, Detlef Weigel8, François Parcy1,2,3,4,12,*, Renaud Dumas1,2,3,4

Affiliations:
1. CNRS, Laboratoire de Physiologie Cellulaire & Végétale, UMR 5168, 38054 Grenoble, France.
2. Univ. Grenoble Alpes, LPCV, F-38054 Grenoble, France.
3. CEA, DSV, iRTSV, LPCV, F-38054 Grenoble, France.
4. INRA, LPCV, F-38054 Grenoble, France.
5. European Molecular Biology Laboratory, 6 rue Jules Horowitz, BP 181, 38042 Grenoble, France.
6. Unit of Virus Host-Cell Interactions, UJF-EMBL-CNRS, UMI 3265, 6 rue Jules Horowitz, 38042 Grenoble Cedex 9, France.
7. Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK.
8. Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany
9. Botanisches Institut, Lehrstuhl I, Universität zu Köln, Biozentrum Köln, Zülpicher Str. 47b, 50674 Köln, Germany
10. BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China
11. Department of Biological Sciences, Department of Medicine, University of Alberta, Edmonton AB, T6G 2E9, Canada
12. Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, University of British Columbia, Vancouver, British Columbia, V5Z 4H4, Canada.
a Present address: Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK.
b Present address: Research School of Biology, The Australian National University, Acton, ACT 0200
‡ These authors made equal contributions
Correspondence to: François Parcy (francois.parcy@cea.fr) or Max Nanao (mnanao@embl.fr)

Abstract: (127 words)

Transcription factors are key players in evolution. Changes affecting their function can yield novel life forms but also have deleterious effects. Consequently, gene duplication events that release one gene copy from selective pressure are thought to be the common mechanism by which transcription factors acquire new activities. Here we show that LEAFY, a major regulator of flower development and cell division in land plants, underwent changes to its DNA binding specificity, even though plant genomes generally contain a single copy of *LEAFY*. We examined how these changes occurred at the structural level, and identify an intermediate LEAFY form in hornworts that appears to adopt all different specificities. This promiscuous intermediate could have smoothed the evolutionary transitions thereby allowing LEAFY to evolve new binding specificities while remaining a single copy gene.

One Sentence Summary:

The single copy, essential and highly conserved LEAFY plant transcription factor evolved DNA binding specificities through a promiscuous intermediate.

Main Text (2524 words including acknowledgments, notes, references and figures captions):

The rewiring of transcriptional networks is an important source of evolutionary novelty (*1*-*3*). Variation often occurs through changes in *cis*- regulatory elements which are DNA sequences containing binding sites for transcription factors (TF) regulating nearby genes (*3*, *4*). There is less evidence for regulatory changes affecting the protein-coding sequence of TFs. Such changes are expected to be under highly stringent selection because they could impair the expression of many downstream targets. Gene duplication provides a solution to this dilemma as additional TF gene copies may acquire new functions, provided that the aggregate copies fulfill the function of the original TF (*5*). Indeed TF DNA binding specificity has been shown to diversify within multigene families (*6, 7*). In some cases, however, TF coding genes remain single copy due to phenomena such as paralog interference (*8*) which can impede neo-functionalization. When essential TFs are maintained as single copy genes, the extent to which they can evolve is not clear. To address this question, we examined the *LEAFY (LFY)* gene as an evolutionary model.

Except in gymnosperms, where two paralogs (*LEAFY* and *NEEDLY*) are usually present (**Fig. 1A)**, *LFY* exists mostly as a single copy gene in land plants (*9*). It plays essential roles as a key regulator of floral identity in angiosperms and of cell division in the moss *Physcomitrella patens* (*10*). *LFY* encodes a transcription factor, which binds DNA through a highly conserved dimeric DNA binding domain (DBD) (*11*). Despite this conservation PpLFY1, a LFY homolog from the moss *P. patens*, is unable to bind the DNA sequence recognized by LFY from *Arabidopsis thaliana* (AtLFY) (9) suggesting that LFY DNA binding specificity might have changed during land plant evolution.

We mined the transcriptomes from algal species, whose origin predates the divergence of mosses and tracheophytes, and found *LFY* homologs in six species of streptophyte green algae (**Figs. 1A, S1**). Thus, *LFY* is not specific to land plants. Despite this extended ancestry, the LFY-DBD sequence, including the amino acids in direct contact with DNA, remains highly conserved (**Figs. 1B, S1**). We used HT-SELEX (*12*) experiments to systematically analyze the DNA binding specificity of LFY proteins from each group of plants. After optimizing alignments (*13*), we found that the SELEX motifs fell into three groups (**Figs. 1C, S2**), suggesting that LFY changed specificity at least twice.

Most LFY proteins from land plants (angiosperms, gymnosperms, ferns and liverworts) bind the same DNA motif (type I) as AtLFY (*13*). PpLFY1, however, binds to a different motif (type II), despite possessing the same 15 DNA binding amino acids as AtLFY (**Fig. 1B)**. These SELEX results explain why all embryophyte LFY homologs, except PpLFY1, display AtLFY-like activity when expressed in *A. thaliana (9)*. Motifs I and II share a similar overall organization, consisting of two 8-bp inverted half-sites separated by 3 nucleotides, but their peripheral positions differ. The newly identified hornwort and algal LFY proteins bind to a third motif (type III) that resembles motif II, but without the central 3-bp spacer **(Fig. 1C)**. With AtLFY, PpLFY1 and KsLFY (from *Klebsormidium subtile*) as representative proteins of the three specificities, we confirmed that each protein displays a strong preference for one motif type (**Figs. 1D, S3, table S1**).

Given the broad conservation of the LFY-DBD sequence, we asked how these different specificities could be explained molecularly. We solved the crystal structure of PpLFY1-DBD bound to a motif II DNA (**Fig. 2A, table S2**) and compared it to the previously determined AtLFY1-DBD dimer/type I DNA complex (*11*). The two ternary complexes are highly similar (RMSD of protein backbone atoms of 0.6 Å). However, PpLFY1-DBD makes additional contact with DNA: aspartic acid 312 (D312) interacts with the cytosine base (C) at position -6 of the DNA binding motif, which is the nucleotide most different between motifs I and II obtained by SELEX **(Figs. 1C, 2B)**. In AtLFY, position 312 is occupied by a histidine residue (H312), which is pulled away from the DNA by an arginine (R345), a conformation that precludes direct H312/DNA contact. In contrast, in PpLFY1, a cysteine residue (C345) replaces R345, which does not affect the positioning of D312, thus allowing it to contact the cytosine base. To test the importance of positions 312 and 345, we swapped these residues between PpLFY1 and AtLFY (**Figs. 2C, D**). This was sufficient to convert specificity from type I to type II and vice-versa, confirming the key role of these two positions. This result is consistent with an *in vivo* study showing that a PpLFY1-D312H mutant can bind a type I sequence and partially complement a *lfy* mutation in *A. thaliana* plants (*9*).

We next investigated binding to motif III. Motif III half-sites are similar to those of motif II (**Fig. 1C**) owing to the presence of a glutamine (Q) at position 312 in type III LFYs: Q is known to interact with multiple bases (*14*) (**fig. S4**) and the small residues present at position 345 (cysteine, alanine or serine) allow Q312 to freely interact with position -6. Critically, motif III differs from motif II by the lack of the central 3-bp spacer (**Fig. 1C**). Modelling a LFY-DBD/ motif III ternary complex by removing the 3-bp spacer in the type II DNA sequence (**Fig. 3A**) revealed that the interaction between helices α1 and α7, which stabilizes dimeric AtLFY- and PpLFY1-DBD positioning (*11*), could no longer exist for motif III.

Consistent with this observation, interacting regions of helices α1 and α7, including the key amino acid H387 on α7 (*11*), are highly conserved from bryophytes to angiosperms (type II and I), but are variable in algae (type III) (**Figs. 1B, S1**). To test the importance of the α1/α7 interaction in binding to 3-bp spaced half-sites, we mutated PpLFY1 H387 and R390 residues (which make most α1/α7 contacts). This was sufficient to shift the DNA binding preference of PpLFY1 from type II to type III (**Fig. 3B).** These observations suggest that LFY-DBD preferentially binds to 3-bp spaced half-sites (motifs I and II) when the α1/α7 interaction surface is present, and to motif III in the absence of this surface. Nevertheless, both the pseudo symmetry of motif III (**fig. S2**) and the size of LFY/DNA complexes (**fig. S4**) suggest that LFY binds motif III as a dimer, possibly through an alternative dimerization surface. These analyses pinpoint the molecular basis of DNA specificity changes to three amino acid sites: positions 312 and 345 determine the half site sequence, and position 387 determines the dimerization mode.

However, if, as shown in *P. patens* and angiosperms, LFY plays a key role throughout plant evolution, how could these changes have been tolerated? Because once arisen, they would have instantaneously modified the expression of the entire set of LFY target genes. Here, our *LFY* phylogeny (**fig. S5**) yields two insights: 1) Although we cannot completely rule out the occurrence of transient ancient duplications, all known duplication events occurred subsequent to changes in the binding specificity of the protein; therefore the *LFY* gene likely evolved new DNA binding modes independently of changes in copy number 2) the hornwort *LFY* lineage diverges from a phylogenetic node that lies between the type III and type I/II binding specificities. On closer examination, we realized that NaLFY from the hornwort *Nothoceros aenigmaticus* had type III specificity according to the SELEX experiment despite having the H387 dimerization residue typical for type I and II specificities (**Fig. 1B, C**). Using Electrophoretic Mobility Shift Assay (EMSA) experiments, we assayed NaLFY and NaLFY-DBD DNA binding, and found that their dimers (**fig. S6**) could bind all three types of DNA motifs (**Figs. 4, S3, S7).** We also established that NaLFY binding to motifs I and II was allowed by the presence of a functional α1/α7 interaction surface (**Fig. 4)**. The SELEX experiment most likely identified only motif III because of its slightly more efficient binding to NaLFY (**fig. S3, Table S1**).

Our amino acid reconstruction analyses across the LFY phylogeny identify the phylogenetic location of the three specificity transitions that occurred during LFY evolution (**Figs. 4, S8**). Initially, the ancestral algal LFY bound motif III as a dimer (with Q312 and C345 half-site determinants). Subsequently the evolution of the α1/α7 interaction surface generated a promiscuous LFY intermediate with two modes of DBD dimerization and a versatile Q at position 312, which bound all three types of DNA motifs. Mutations affecting positions 312 and 345 then completed the transition to type I or II specificities. While this precise path cannot be unambiguously determined by reconstruction alone (**Figs. 4, S8**), the biochemical data reveals that two LFY states (Q312-C345 and H312-C345) bind to both motifs I and II (**Figs. 2C, 4**), Our scenario, using either of these two states as an intermediate, provides an evolutionary route through a promiscuous platform that avoids deleterious transitions. Furthermore, this scenario is equally parsimonious in the context of all alternative organismal phylogenetic hypotheses (**fig. S9**). Whether these transitions were accompanied by a complete change in target gene sets or whether some *cis*-elements coevolved with DNA binding specificity (*15*) is unknown. Scanning the *P. patens* genome for PpLFY1 binding sites does not suggest any global conservation of targets but identified several MADS-box genes potentially bound by LFY in both *Arabidopsis* and *P. patens* **(Table S3)**.

In conclusion, a highly conserved and essential TF evolved radical shifts in DNA binding specificity by a mechanism that does not require gene duplication. Detailed structural characterization of the different modes of DNA binding across the transition to land plants enabled us to capture LFY in a state of increased promiscuity that has persisted in *N. aenigmaticus*. This promiscuous intermediate likely facilitated the evolutionary transition between specificities as previously shown for the evolution of metabolic enzymes or nuclear receptors (*16-18*). While we have focused on the more intractable problem of evolution in single copy TFs, it is plausible that the mechanisms we describe could also contribute to the evolution of TFs encoded by multigene families.

References and Notes:

1. S. B. Carroll, *Cell* **101**, 577 (2000).
2. I. S. Peter, E. H. Davidson, *Cell* **144**, 970 (2011).
3. B. Prud'homme, N. Gompel, S. B. Carroll, *Proc. Natl. Acad. Sci. U.S.A.* **104 Suppl 1**, 8605 (2007).
4. G. A. Wray, *Nature Rev. Genet.* **8**, 206 (2007).
5. H. E. Hoekstra, J. A. Coyne, *Evolution* **61**, 995 (2007).
6. G. Badis *et al.*, *Science* **324**, 1720 (2009).
7. M. F. Berger *et al.*, *Cell* **133**, 1266 (2008).
8. C. R. Baker, V. Hanson-Smith, A. D. Johnson, *Science* **342**, 104 (2013).
9. A. Maizel *et al.*, *Science* **308**, 260 (2005).
10. T. Tanahashi, N. Sumikawa, M. Kato, M. Hasebe, *Development* **132**, 1727 (2005).
11. C. Hamès *et al.*, *EMBO J* **27**, 2628 (2008).
12. Y. Zhao, D. Granas, G. D. Stormo, *PLoS Comput. Biol.* **5**, e1000590 (2009).
13. E. Moyroud *et al.*, *Plant Cell* **23**, 1293 (2011).
14. N. M. Luscombe, R. A. Laskowski, J. M. Thornton, *Nucleic Acids Res.* **29**, 2860 (2001).
15. C. R. Baker, B. B. Tuch, A. D. Johnson, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7493 (2011).
16. O. Khersonsky, C. Roodveldt, D. S. Tawfik, *Curr. Opin. Chem. Biol.* **10**, 498 (2006).
17. A. Aharoni *et al.*, *Nat. Genet.* **37**, 73 (2005).
18. G. N. Eick, J. K. Colucci, M. J. Harms, E. A. Ortlund, J. W. Thornton, *PLoS Genet.* **8**, e1003072 (2012).
19. A. Dummler, A. M. Lawrence, A. de Marco, *Microb. Cell Fact.* **4**, 34 (2005).
20. D. M. Goodstein *et al.*, *Nucleic Acids Res.* **40**, D1178 (2012).
21. D. J. Zwickl, The University of Texas at Austin (2006).
22. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* **19**, 1572 (2003).
23. W. P. Maddison, R. Maddison. D., in http://mesquiteproject.org. (2011).
24. J. P. Bollback, *Bmc Bioinformatics* **7**, (2006).
25. H. Chahtane *et al.*, *Plant J.*, (2013).
26. T. L. Bailey, C. Elkan, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28 (1994).
27. D. Flot *et al.*, *J. Synchrotron Radiat.* **17**, 107 (2010).
28. A. J. McCoy *et al.*, *J. Appl. Crystallogr.* **40**, 658 (2007).
29. P. Emsley, K. Cowtan, *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126 (2004).
30. G. Bricogne, *Acta Crystallogr. D Biol. Crystallogr.* **49**, 37 (1993).
31. E. I. Barker, N. W. Ashton, *Plant Cell Rep* **32**, 1161 (2013).
32. A. C. Cheng, W. W. Chen, C. N. Fuhrmann, A. D. Frankel, *J. Mol. Biol.* **327**, 781 (2003).
33. N. M. Luscombe, R. A. Laskowski, J. M. Thornton, *Nucleic Acids Res.* **29**, 2860 (2001).
34. O. Malek, K. Lattig, R. Hiesel, A. Brennicke, V. Knoop, *Embo J.* **15**, 1403 (1996).
35. Y. Chang, S. W. Graham, *Am. J. Bot.* **98**, 839 (2011).
36. Y. L. Qiu, Y. R. Cho, J. C. Cox, J. D. Palmer, *Nature* **394**, 671 (1998).
37. Y. L. Qiu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 15511 (2006).
38. Y. L. Qiu *et al.*, *Int. J. Plant Sci.* **168**, 691 (2007).
39. D. L. Nickrent, C. L. Parkinson, J. D. Palmer, R. J. Duff, *Mol. Biol. Evol.* **17**, 1885 (2000).
40. V. V. Goremykin, F. H. Hellwig, *Plant Systematics and Evolution* **254**, 93 (2005).
41. T. Nishiyama *et al.*, *Mol. Biol. Evol.* **21**, 1813 (2004).
42. L. Gao, Y. J. Su, T. Wang, *J. Syst. Evol.* **48**, 77 (2010).

43. E. Moyroud, E. Kusters, M. Monniaux, R. Koes, F. Parcy, *Trends Plant Sci* **15**, 346 (2010).

Acknowledgements: We thank M. Schmid for help with sequencing, E. Masson, J. Kyozuka, M. Hasebe, C. Scutt and C. Finet, J.C. Villarreal for sharing materials, A. Maizel and S. Rensing for discussion, A. Maizel, R. Worsley-Hunt, A. Mathelier, M. Blazquez, O. Nilsson, C. Zubieta, and J. Chen for critical reading of the manuscript. This work was supported by funds from the Max Planck Society to D.W., the ANR (SVSE2–2011) and Rhône-Alpes CMIRA support to F.P. and the FP7 P-CUBE n° 227764 to R.D. The 1000 Plants (1KP) initiative, led by G.K.S.W., is funded by Alberta EAE/AITF/iCORE, Musea Ventures, and BGI-Shenzhen. Data deposition: Genbank accessions (*AmboLFY*: KF193872; *NaLFY*: KF269532; *KsLFY*: KF269535; *CsLFY*: KF269533; *CvLFY*: KF269536; *CyLFY*: KF269534) and Protein Data bank accession 4BHK for the reported crystal structure. The authors declare no competing financial interests.

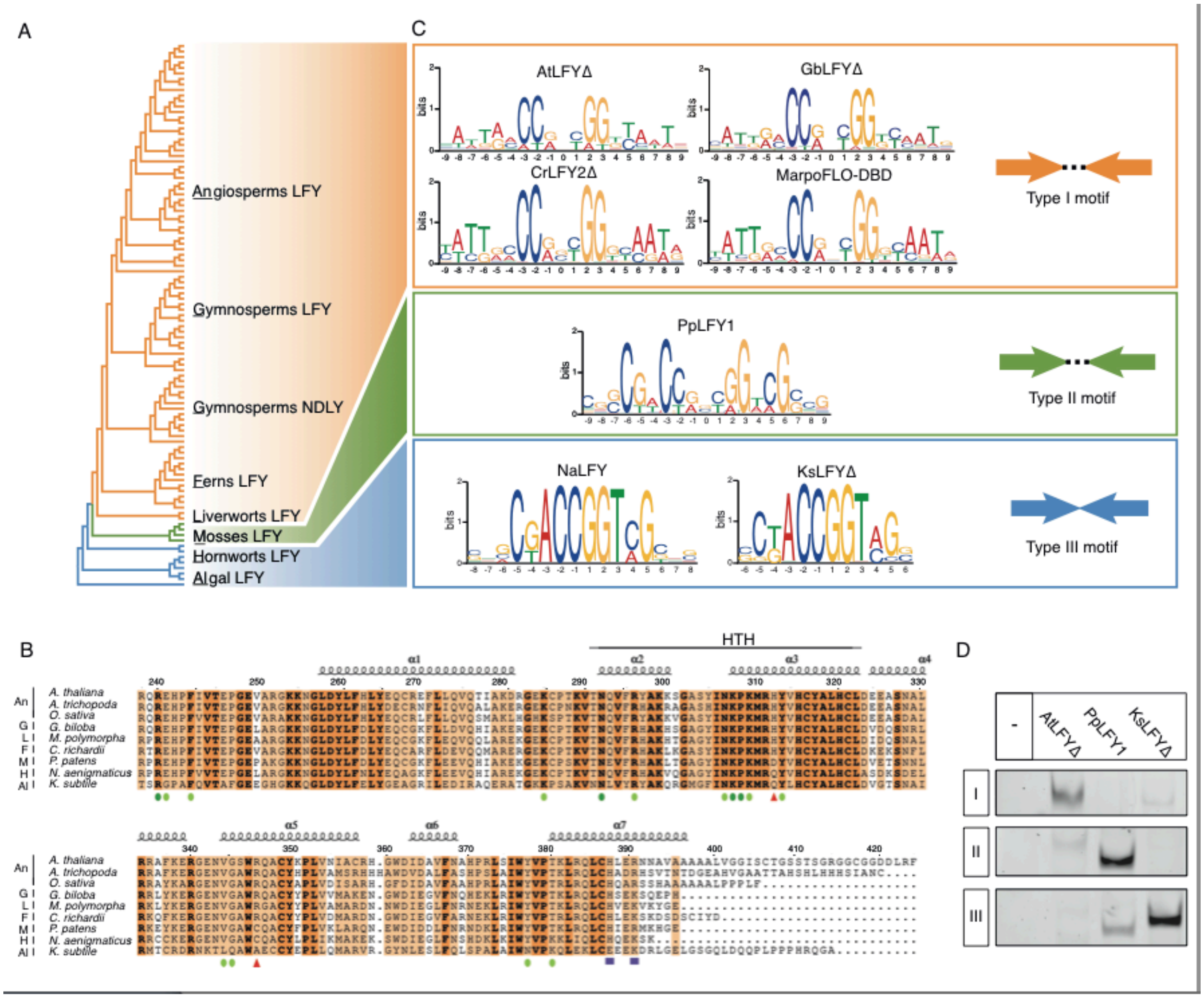

Fig. 1. Evolution of LFY DNA binding specificity. (A) Simplified *LEAFY* phylogeny (detailed in fig. S5). DNA binding specificities are color coded: type I (orange), II (green) or III (blue). (**B**) Alignment of LFY-DBDs. Amino acids numbering and secondary structure annotation (α, alpha helices, HTH, helix-turn-helix domain) are based on AtLFY from *Arabidopsis thaliana*. Dark green dots: DNA base contacts, light green dots: phosphate backbone contacts, red triangles: residues involved in the PpLFY1-specific DNA contacts, purple rectangles: residues involved in the interaction between DBD monomers. (**C**) SELEX motifs for AtLFYΔ, GbLFYΔ (*Ginkgo biloba*), CrLFY2Δ (*Ceratopteris richardii*), MarpoFLO (*Marchantia polymorpha*), PpLFY1 (*Physcomitrella patens*), NaLFY (*Nothoceros aenigmaticus*), KsLFYΔ (*Klebsormidium subtile*), Δ: proteins starting at amino-acid 40 (on the basis of AtLFY sequence). Cartoons depict binding site organization: half site (arrow) with or without a 3-bp spacer. (**D**) EMSA with AtLFYΔ, PpLFY1 and KsLFYΔ proteins (10 nM) and the three types (I, II, III) of DNA probes. Only the protein-DNA complexes are shown.

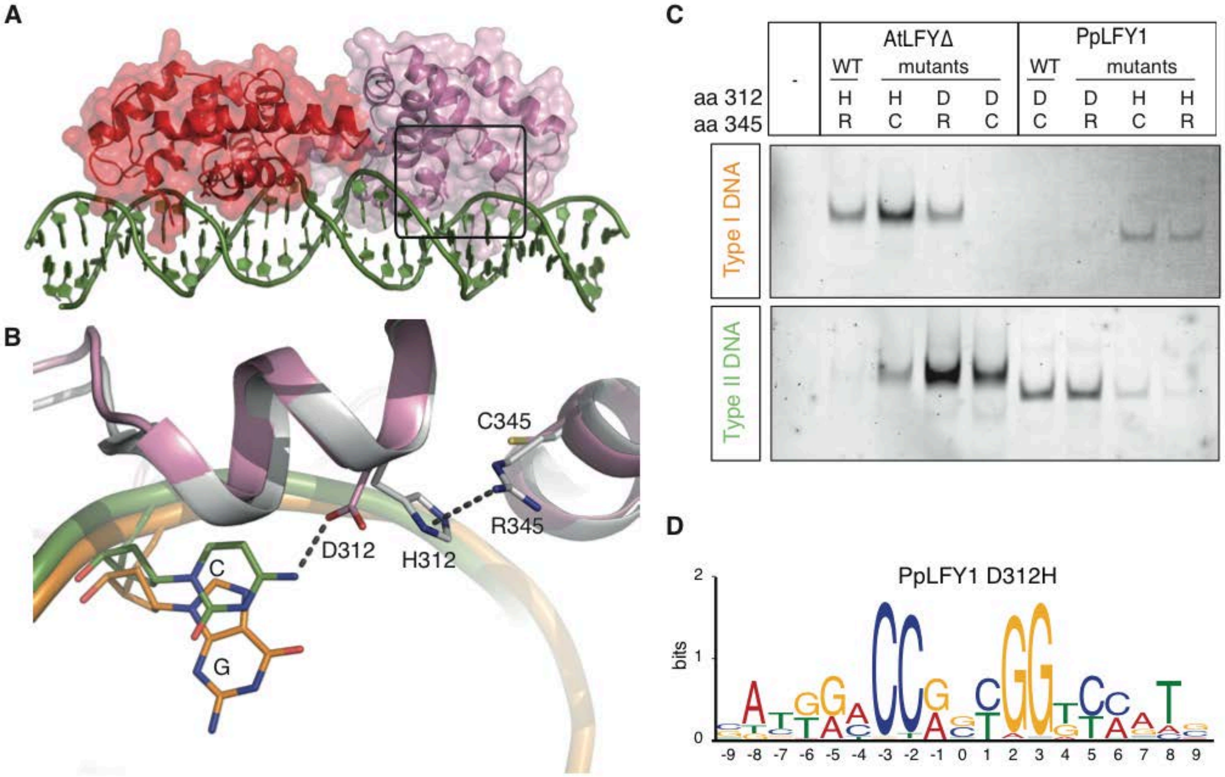

Fig. 2. Structural basis for type II specificity. (**A)** Crystal structure of PpLFY1-DBD (red and pink) bound to DNA (green). The squared area is detailed in **(B)** after applying a 70° rotation. (**B**) Superimposition of AtLFY-DBD (grey) **-** DNA (orange) and PpLFY1-DBD (pink) **-** DNA (green) complexes. Specificity determinant residues and bases are represented as sticks. For amino acids, H: histidine R: arginine, D: aspartate, C: cysteine. For DNA bases, C: cytosine, G: guanine. (**C**) Effect of specific mutations on the DNA binding specificity of AtLFYΔ and PpLFY1 in EMSA. Note that the H312, C345 combination allow binding to both motifs I and II. All proteins are at 25 nM and only the protein-DNA complexes are shown. (**D**) SELEX motif of the PpLFY1- D312H protein, bearing strong resemblance to motif I.

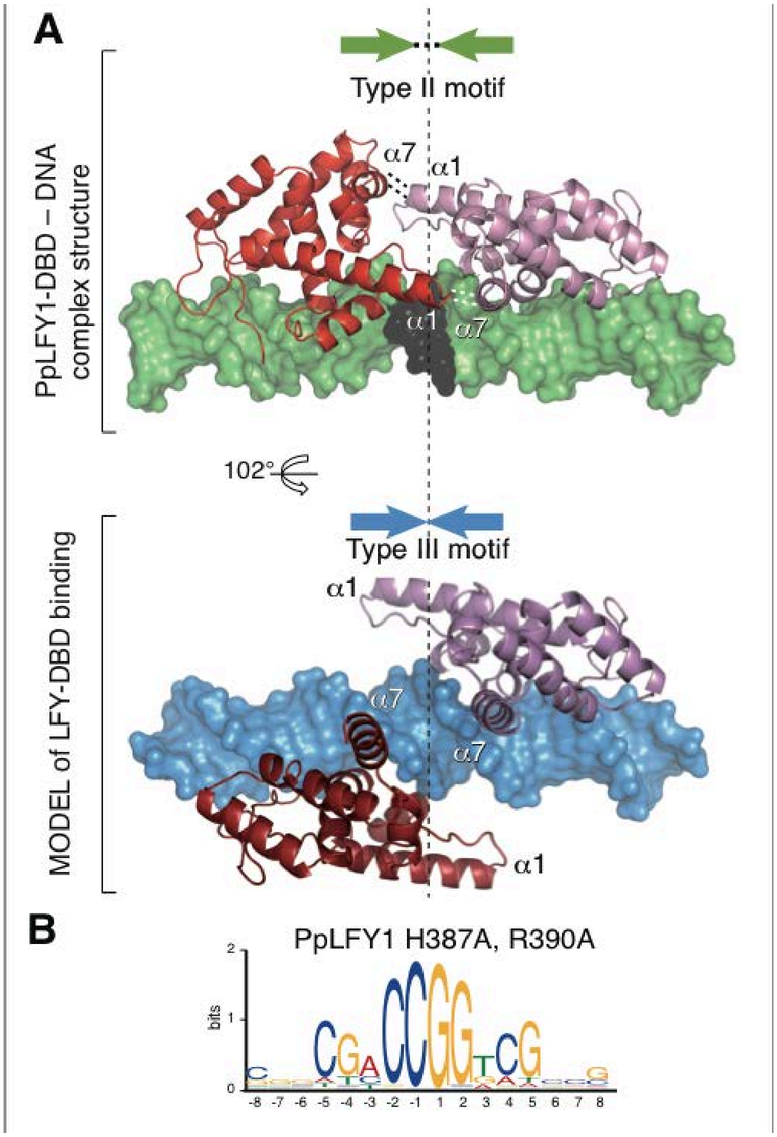

Fig. 3. Structural model for type III specificity. (**A**) Top: PpLFY1-DBD dimer (in red and pink) bound to DNA (in green except the black 3-bp spacer). Interaction between monomers (involving alpha helices α1 and α7) are shown with dashed lines. Bottom: Modelled type III binding with DNA in blue. (**B**) SELEX motif of PpLFY1-H387A, R390A, showing strong resemblance to motif III.

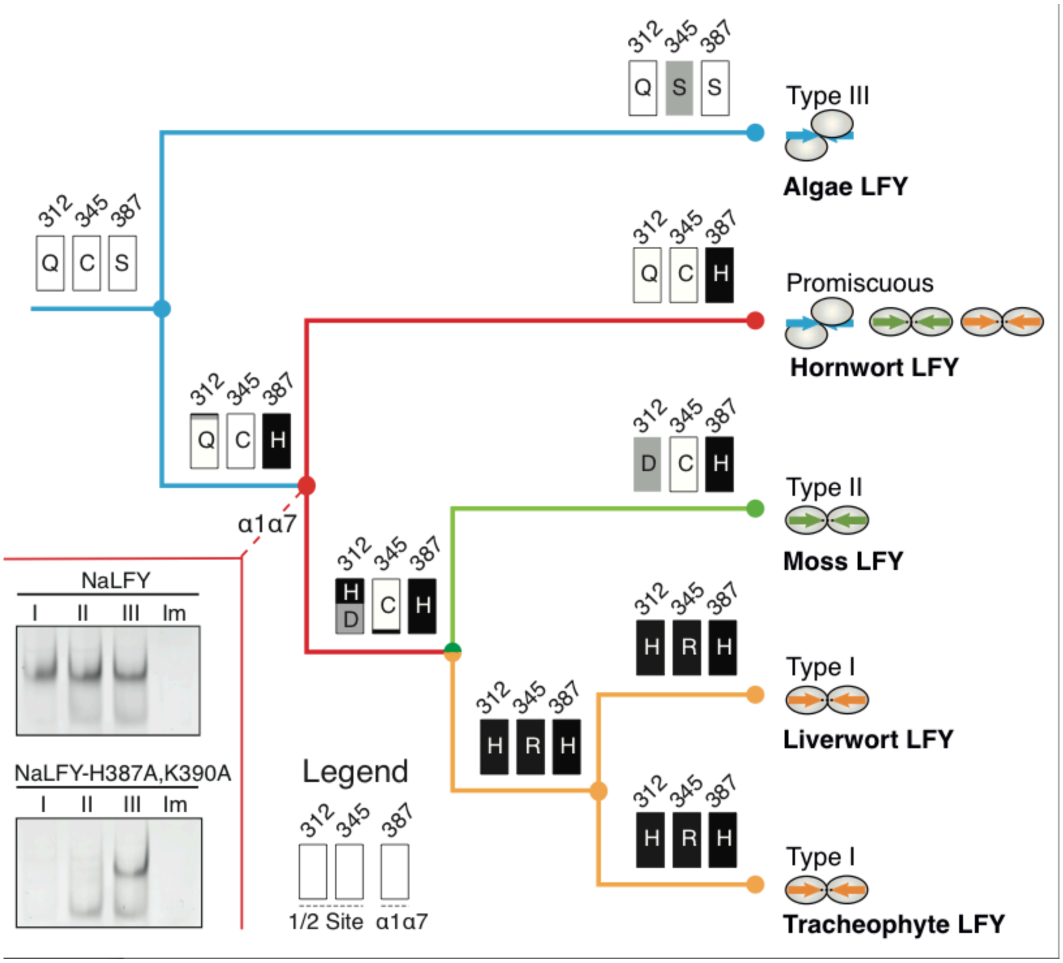

Fig. 4. **Proposed evolution of LFY DNA binding specificity in green plants.** The bayesian estimation of the posterior probability of ancestral states for amino-acid positions 312, 345 and 387 is depicted at the major phylogenetic nodes. Probabilities for different residues at a given position and node are indicated by the relative size of stacked boxes. The analysis shows that the ancestral LFY most probably possessed a type III specificity, and that the promiscuous form arose when land plants emerged. DNA binding specificity is color-coded (type I: orange, type II: green, type III: blue and relaxed specificity: red). 1 α1/α7 refers to the α1/α7 dimerization interface. Inset: NaLFY interacts with all three types of DNA binding motifs in EMSA (see also fig. S7), but not with the type I mutated probe (Im). The H387A, K390A mutations reduced the binding to type I or II motifs, but not to type III. Both proteins are at 1 µM, only the protein-DNA complexes are shown.