**Updated phylogeny and protein structure predictions revise the hypothesis on the origin of MADS-box transcription factors in land plants**

Yichun Qiu [1,2,†], Zhen Li [3,4,†], Dirk Walther [1], Claudia Köhler [1,2,*]

[1] Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany, 14476.

[2] Swedish University of Agricultural Sciences & Linnean Center for Plant Biology, Uppsala BioCenter, Uppsala, Sweden, 75007.

[3] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, 9052.

[4] VIB Center for Plant Systems Biology, VIB, Ghent, Belgium, 9052.

[†] equal contribution

[*] corresponding author: Claudia Köhler

Mail: Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, Potsdam-Golm, Germany, 14476

Phone: +49 331 567-8100

Email: Koehler@mpimp-golm.mpg.de

**Key words**: MADS-box transcription factors, land plants, MEF2, M-type, gene duplication

**Abstract**

MADS-box transcription factors (TFs) are broadly present in eukaryotic genomes. Varying by domain structures, MADS-box TFs in plants are categorized into M-type and MIKC-type. For about twenty years, M-type genes were considered closely related to the SRF genes in animals, collectively referred to as Type I MADS-box genes. MIKC-type and animal MEF2 genes were grouped as Type II, presumably duplicated with Type I genes before the divergence of eukaryotes. Exploiting available genomic data, we reassessed the evolutionary history of eukaryotic MADS-box genes and propose an alternative hypothesis. Our phylogenetic analyses support the ancient duplication of SRF/MEF2; however, both M-type and MIKC-type originated from the lineage of MEF2 via another duplication event before the divergence of land plants. Protein structures predicted by AlphaFold2 support this evolutionary scenario, with both M-type and MIKC-type proteins in plants resembling the MEF2 3D structure, distinct from SRF. Therefore, we propose that the most recent common ancestor of Archaeplastida (the kingdom Plantae) likely did not inherit any SRF gene. The retained MEF2 TFs acquired a Keratin-like domain and became MIKC-type upon the evolution of Streptophyta. Subsequently in land plants, M-type TFs evolved from a duplicated MIKC-type precursor through loss of the Keratin-like domain. M-type and MIKC-type then largely expanded and functionally differentiated in concert with the increasing complexity of land plant body architecture. We attribute the adaption to the terrestrial environment partly to the divergence among MEF2-type MADS-box genes and the repetitive recruitment of these originally stress-responsive TFs into developmental programs, especially those underlying reproduction.

## Introduction

MADS-box transcription factors (TFs) are broadly present in eukaryotes. They regulate diverse and important biological functions as reported in animals, fungi, plants, and protists (reviewed in Messenguy and Dubois 2003). The name derives from the four founding members, Minichromosome maintenance 1 (Mcm1) from *Saccharomyces cerevisiae*, AGAMOUS from *Arabidopsis thaliana*, DEFICIENS from *Antirrhinum majus*, and Serum response factor (SRF) from *Homo sapiens* (Schwarz-Sommer et al. 1990). Animal genomes generally encode for two types of MADS-box genes, SRF and myocyte enhancer factor-2 (MEF2) genes that are present in one to a few copies. The budding yeast *Saccharomyces cerevisiae* has four MADS-box genes; Mcm1 and Arg80 are related to the animal SRF, and Rlm1 and Smp1 are related to MEF2. Several phylogenetic analyses inferred the origin of SRF and MEF2 types through an ancient gene duplication event (Fig.1a,b; Theissen et al. 1996; Alvarez-Buylla et al. 2000; Gramzow et al. 2010). After the gene duplication the two ancestral genes of the distinct types diverged by the domains downstream of the MADS domain; while SRF type TFs are characterized by a SAM domain (SRF, ARG80 and MCM1), the corresponding region in MEF2 type TFs is referred to as the MEF2 domain (Shore and Sharrocks 1995). The crystal structures of several MADS-box TFs have been resolved, including human SRF and budding yeast MCM1, human MEF2A and mouse MEF2C. The conserved MADS domain comprises an alpha helix, followed by two antiparallel beta strands. MEF2 and SRF structures differ in the second alpha helix, constituted by the SAM or MEF2 domain, respectively, where a kink in the SAM domain of SRF TFs changes the orientation of the second helix in the opposite direction to that of MEF2 TFs (Fig.2a). The two alpha helices and the connecting beta strands build up the interface for TF dimerization and DNA-binding.

In land plants, MADS-box TFs (generally known as AGAMOUS-like (AGL) genes) have evolved to be a flourishing family with typically as many as 50 to over 100 members in angiosperms, in sharp contrast to the family size of MADS-box genes in other eukaryotes (Gramzow and Theissen 2013). According to the domain architecture, the MADS-box TFs in land plants can be specified into two types: while the M-type (Type I) TFs share no well-

3

known conserved domain following the MADS-box domain, the MIKC-type (Type II) TFs have the MADS-box domain followed by the Intervening, Keratin-like and C-terminal domains, among which the K domain is likely specific for plants (Alvarez-Buylla et al. 2000; Parenicová et al. 2003). Due to their critical roles in establishing floral organ identity in angiosperms, the MIKC-type MADS-box genes have been extensively studied. In contrast, M-type MADS-box genes have only been identified along with the first angiosperm genome of *Arabidopsis thaliana*, and gradually emerging studies have linked their functions to the development of the female gametophyte and endosperm (Bemer et al. 2010; Masiero et al. 2011; Qiu and Kohler 2022). Besides their distinct domain arrangements, land plants M-type and MIKC-type MADS-box genes vary in their expression domains, numbers of exons; and noticeably, M-type MADS-box genes are fast evolving and frequently undergo duplication and loss (Parenicová et al. 2003; Nam et al. 2004).

Among many known regulatory functions, MADS-box TFs act as major regulators of plant reproduction and have been closely connected with the rise of flowering plants to ecological dominance (Ng and Yanofsky 2001; Kaufmann et al. 2005). The family size of MADS-box genes has been linked to the complexity of the plant body plan (Theissen et al. 1996; Thangavel and Nayar 2018; Kaufmann et al. 2005). Thus, solving the origin and subsequent diversification of plant MADS-box genes is required to understand the evolutionary success of land plants.

Upon the discovery of M-type MADS-box genes, a timely survey suggested that M-type and MIKC-type genes in plants are orthologous to the SRF and MEF2 genes in other eukaryotic lineages, respectively (Alvarez-Buylla et al. 2000). Based on this model, an ancient duplication before the divergence of the extant eukaryotic lineages gave birth to the two classes of MADS-box genes in plants, likewise in animals and other eukaryotes. This model has been influential in the field of MADS-box evolution and served as a basis for investigations of MADS-box gene evolution across all phylogenetic scales. Nevertheless, some thoughtful critics on this model have been neglected for a while (Kaufmann et al. 2005; De Bolt et al. 2003). Limited by the data availability twenty years

ago, indeed, as noted by the authors (Alvarez-Buylla et al. 2000), the clustering of M-type TFs in plants and SRF TFs in animals and fungi was not well supported in the original study based on only a few Arabidopsis, animal, and fungal sequences.

The emerging genomes of Charophytes, the paraphyletic algal relatives of land plants, have shed light on the evolution of gene families underlying the successful terrestrialization of land plants. In the genomes of *Chara braunii* and *Klebsormidium flaccidum*, only MIKC-type MADS-box genes are present (Nishiyama et al. 2018; Hori et al. 2014). Furthermore, in the genomes of the green algae *Chlamydomonas reinhardtii*, *Ostreococcus tauri*, *Ostreococcus lucimarinus* and the red algae *Cyanidioschyzon merolae*, no MIKC-type gene could be detected; however, the annotated MADS-box TFs, although without a K domain, are identified as MEF2 type (Type II) (Kaufmann et al., 2005; Thangavel and Nayar 2018). Thus, SRF type MADS-box genes have so far never been found in the charophycean, green or red algae. If M-type MADS-box TFs in land plants are descendants of ancestor SRF genes as stated by the popular model, the orthologous SRF genes should have been lost convergently and repeatedly in all of successive sister groups. Thus, the systematic lack of SRF type genes in these lineages challenges the orthology between M-type MADS-box TFs in land plants and SRF TFs in animals and fungi. We therefore propose an alternative model explaining the evolution of MADS-box genes in the green lineages.

**Results**

**Ancient duplication of MEF2 and SRF clades is supported using newly available genome sequences**

Capitalizing on newly available genome sequences spanning the broad phylogeny of eukaryotes, we re-evaluated the original phylogenetic model of MADS-box gene evolution with extended sample sequences. We searched published genomes of different lineages of eukaryotes and identified MADS-box genes in 175 species (Supplementary table 1). The species represent seven currently accepted eukaryotic groups: 1.

5

Archaeplastida (e.g., streptophytes including land plants, green algae, Prasinodermophyta, red algae, and glaucophytes), 2. Cryptista, 3. Haptista, 4. SAR supergroup which are Stramenopila (e.g., brown algae, diatoms, oomycetes) / Alveolata (e.g., ciliates, dinoflagellates, Apicomplexa) / Rhizaria, 5. Amorphea (e.g., animals, fungi, amoebae), 6. Discoba, and 7. Metamonada (the latter two formerly were collected in Excavata) (Burki et al. 2020).

Using the MADS-box domain sequences as input, we inferred phylogenetic trees with maximum likelihood, neighbour joining, and Bayesian inference. All three methods consistently find that the MADS domain sequences naturally form two major clades, corresponding to current SRF and MEF2 lineages, as referenced by the known SRF and MEF2 genes from animals, fungi, and amoebae (Fig.3; Supplementary Fig.S1). This finding supports the overarching hypothesis that an ancient duplication of a MADS-box gene gave rise to the SRF and MEF2 precursors. We found both SRF and MEF2 types of MADS-box genes in nearly all of the surveyed Amorphea species, as well as Cryptista and Haptista. Furthermore, two species in Discoba, representing a distinct group distantly diverged from the plant and animal lineages, have both types of MADS-box genes (Supplementary table 1). Together, this data provides supporting evidence for the presence of SRF and MEF2 type MADS-box genes early before the diversification of modern eukaryotes.

**MEF2 and SRF MADS-domains form distinct structures**

SRF and MEF2 type MADS-box domains are known to form distinct protein structures (Fig.2a). We applied AlphaFold2 to predict the structures of identified eukaryotic MADS-box TFs and categorized predicted structures as either SRF or MEF2 type based on their overlay pattern with resolved MADS-box structures (Fig.2b,c; Supplementary table 2). We first analyzed the structures of Amorphea and Discoba TFs that inferred by phylogeny belong to the MEF2 type. Indeed, the human SRF model was rejected, while the human MEF2A model fitted well. In contrast, the sequences present in the SRF clade fitted significantly better the human SRF model than the human MEF2A model (Fig.2b,c;

6

Supplementary Fig.S2). This data indicates that the predicted structure of the MADS-domain aids to classify proteins into MEF2 or SRF.

**Land plant Type I and II MADS-box TFs are both MEF2 type**

Previous analyses predicted that Type I (M-type) MADS-box genes in land plants are more closely related to the SRF type (Alvarez-Buylla et al. 2000). In contrast, our phylogenetic analyses using three different approaches consistently predicted that Type I MADS-box genes in land plants, clustered within the MEF2 clade. The latter, as expected, also includes plant Type II (MIKC-type) genes. Thus, both M-type and MIKC-type genes are inferred to be MEF2 type and no SRF gene is present in land plants. We further tested this finding using structural analyses. With the predicted models by AlphaFold2, MIKC-type TFs in land plants resemble the MEF2 structure (Fig.2d). Similarly, M-type TFs in land plants all mapped better to the human MEF2A compared to the SRF structure. The second alpha helix of M-type TFs was not predicted to be twisted by a kink, as found in SRF type TFs. Arabidopsis SEPALLATA 3 (AtSEP3) is thus far the only resolved crystal structure of a plant MADS-box protein and as expected, it displays an MEF2 structure as a MIKC-type TF. There is no resolved structure for plant M-type MADS-box TFs up to now. The predicted models for both, MIKC-type) and M-type TFs in land plants, did well align with the AtSEP3 structure (Fig.2d), though with higher alignment scores for MIKC-type TFs, consistent with them being more closely related to AtSEP3. Tightly adjacent to the MADS domain at the C-region, the I domain of MIKC-type TFs resembles the second helix of the MEF2 domain in Amorphea MEF2 type TFs. The second alpha helix in the predicted structural models of the M-type TFs is formed by an Intervening domain-like region (Lai et al. 2021). While initially not defined, the I-like domain in the Arabidopsis M-type TFs has been shown to be required for both dimerization and DNA binding, functionally equivalent to the Intervening domain in MIKC-type TFs (Lai et al. 2021).

**Loss of SRF type genes in the most recent common ancestor of Archaeplastida**

The origin and divergence of M-type and MIKC-type MADS-box genes in land plants can be further inferred from the sister green lineages. In line with previous findings, the

7

presence of only MEF2 type genes and the absence of SRF is observed in the genomes of streptophytic algae, a series of successive sister groups of land plants. *Klebsormidium nitens*, *Mesostigma viride* and *Penium margaritaceum* each have only one MEF2 gene; *Chlorokybus atmophyticus*, *Chara braunii* and *Spirogloea muscicola* have more than one copy, probably because of lineage-specific duplication (Supplementary table 1). The single MEF2 clade of MADS-box genes along the evolution of streptophytes further suggests that M-type and MIKC-type MADS-box TFs were formed by gene duplication in the common ancestor of land plants. The loss of SRF type genes may be tracked back before the diversification of the whole Archaeplastida clade. In green algae and a third lineage of green plants, Prasinodermophyta, represented by *Prasinoderma coloniale*, there is no confidently predicted SRF gene either. In a few species belonging to the core Chlorophyta we found some genes harboring a partial SRF-like MADS-box domain; but these SRF-like genes are quite divergent from other SRF sequences, represented by long branches, low support scores and inconsistent phylogenetic clustering (Supplementary Fig.S3). Their best BLASTP hits against all other MADS-box sequences were fungal SRF genes, indicating that they probably arose from a horizontally transferred fragment in a common ancestor of core chlorophytes. Sister to the green plants, likewise, none of the red algae surveyed has an SRF gene, neither does the glaucophyte *Cyanophora paradoxa*. The lack of SRF genes in the Archaeplastida lineage is unlikely a consequence of multiple independent losses; instead, it raises the more parsimonious model that a common ancestor of the Archaeplastida lineage did not inherit an SRF gene. MEF2 type ancestral genes remained single-copy in many genomes, as reflected by most of the green algae and the glaucophyte algae. However, in a few lineages MEF2 genes got duplicated, for example in some red algae and streptophytes, and most pronouncedly in the land plant lineages. In agreement with this evolutionary scenario, Archaeplastida MADS-box TFs with or without a K domain, were all predicted by AlphaFold2 to align to the MEF2, not the SRF model (Fig.2a; Supplementary Fig.S2). All these MEF2 type structural models share the Intervening or MEF2 domain-like region, which constitutes the second helix with no kink. The same prediction holds, specifically, for MIKC-type TFs in streptophytic algae, which

are close sister genes to the land plant M-type and MIKC-type MADS-box genes (Fig.2a; Supplementary Fig.S2) Thus, the predicted protein structures nicely mirror the suggested phylogeny and support the MEF2 origin of both M-type and MIKC-type MADS-box genes in land plants through lineage-specific gene duplication (Fig.1c).

**Evolution of IKC domains after the MADS domain**

In contrast to the previous proposition that the I domain in plant MIKC-type TFs and the MEF2 domains in animals and fungi were independently acquired, the second alpha helix in all MEF2 type TFs is likely to have a common origin. While previously unrecognized, the I-like region in the M-type TFs and the corresponding region constituting the second helix in the algal Archaeplastida MADS-box TFs are homologous to the I/MEF2 domain and functionally conserved. Meanwhile, the SAM domain present in SRF type TFs has gradually diverged from the precursor of the MEF2 domain, configuring the kink that changes the orientation of the helix. Thus, the MADS-box domain that originally only included the first helix and the antiparallel strands, can be extended to include the second helix, since the helix-strand-helix structure is one functional unit and likely evolved together. The phylogenetic trees generated with the alignments of only the conventional MADS-box region largely agree with the phylogeny inferred from the alignments using the extended region (Fig.3; Supplementary Fig.4).

Analysing the C-terminal sequences of these MEF2 type MADS-box genes in Archaeplastida, we could also infer the gain and loss of the Keratin-like domain. Our data confirmed the proposed streptophytic origin of MIKC-type MADS-box genes (Kaufmann et al., 2005; Thangavel and Nayar 2018) and suggest that a MEF2 gene in ancestral Streptophyta acquired the K domain and continued evolving as the plant-specific MIKC-type. This is supported by the fact that K domains can only be identified in streptophytic MADS-box genes by conserved domain search, but are absent in green algae or Prasinodermophyta, and other eukaryotic groups. Subsequently, upon the rise of M-type genes in land plants by duplication of a MIKC-type ancestral gene, the K domain got lost.

**Sporadic losses of MADS-box TFs across eukaryotes**

Except for the Archaeplastida, some other species have only either SRF or MEF2 type TFs (Supplementary table 1). For example, three species in the Microsporidia, a group of unicellular parasites closely related to fungi, have only SRF type TFs. *Sphaeroforma arctica* and *Thecamonas trahens*, two successive sister groups of animals and fungi, also only have SRF type TFs. In the Haptista, the species *Chrysochromulina tobin* has only the SRF type, but three other species have only MEF2 type TFs, suggesting reciprocal losses after the divergence from a common ancestor comprising two types of MADS-box genes. Species in the SAR group, ciliates, oomycetes and the cercozoan *Plasmodiophora brassicae* have only MEF2 type TFs. Some surveyed species belonging to the green algae, the brown algae, diatoms, dinoflagellates and several Metamonada and Discoba protists among others have no extant MADS-box genes (Supplementary table 1). We tested the possibility that the observed lack of a certain type, or both, of MADS-box TFs is a result of incomplete gene annotations. We therefore scanned the genomes with the profile hidden Markov model for known MADS-box domains (PF00319) from Pfam (Paysan-Lafosse et al. 2022). We indeed identified some unannotated and incomplete MADS-domains in a few species (Supplementary table 3); importantly however, no SRF open reading frame was detected in the Archaeplastida. The most parsimonious explanation for this finding is that SRF genes were lost in the common ancestral lineage leading to the Archaeplastida.

**Discussion**

Supported by an updated phylogeny and models of protein structure, we propose that in the land plant lineage M-type (previously termed Type I and SRF-type) MADS-box TFs arose as a second clade of MEF2 type TFs. Most likely, this was the result of a gene duplication event of a Type II-like (MIKC-type) ancestral gene that was followed by rounds of gene duplication events largely expanding the MADS-box gene family. This new model of the MEF2 evolutionary trajectory is favored by the parsimonious principle considering the absence of SRF genes in the sister lineages of Archaeplastida, specifically those of streptophytic algae (Fig.1c).

The major difference between our proposed model and the previous popular model is the origin of plant M-type MADS-box genes. M-type genes are known to have high substitution rates (Nam et al. 2004), which is reflected by long branches (Fig. 3; Supplementary Fig.1). Previous studies claiming M-type genes in Arabidopsis are less closely related to fungal and animal MEF2 genes have suffered from biased and inadequate sampling. The limitation in sequence sampling also affected another early model, which suggested that M-type genes are polyphyletic, while MIKC-type genes are MEF2-like genes (Kofuji et al. 2003). The surge of eukaryotic genomes filled the gaps between distantly related animal and plant sequences and made it possible to calibrate the incomplete phylogeny of the MADS-box gene family. Specifically, those species representing diverse, previously underrepresented protist groups provided comprehensive support for the hypothesized ancient duplication of SRF and MEF2 types. At the same time, the Charophytic genomes serve as a great reference for the gene family evolution in land plants. We thus took the opportunity to revisit the phylogeny of this gene family that was likely a key driver for plants adapting to land ecosystems.

Our study also confirmed the sharp contrast between MADS-box gene family size in land plants compared to that in other eukaryotes, as previously noted in animals and fungi (Theissen et al. 1996; Thangavel and Nayar 2018). Most eukaryotes have only few MADS-box genes (Supplementary table 1), revealing that the low-copy status remained constant during the evolution of protist-like stages, and including early Archaeplastida. However, following the inferred duplication of M-type and MIKC-type coupled with the terrestrialization of plants, the MADS-box gene family largely expanded, which provided the raw genetic material for subsequent functional differentiation. There have been extensive studies showing that MADS-box TFs are key regulators of plant organ formation, similar to homeobox genes in animals (Nam et al. 2003). The expansion of the MADS-box gene family has been proposed to be linked to the increasing complexity of extant land plants (Gramzow et al. 2014; Thangavel and Nayar 2018). Convergently and in concert with the evolution of multicellularity, while less abundant in copy number, SRF and MEF2 genes in metazoan animals are both functioning in embryo patterning and continue to

regulate muscle development after maturity (reviewed in Potthoff and Olson 2007). Nevertheless, since multicellularity evolved independently in animals and plants, the ancestral function of MADS-box genes may have been different. The missing link for inferring the ancestral functional role of MADS-box genes lies in the unicellular, or under-differentiated multicellular eukaryotes.

Both MEF2 and SRF genes in unicellular and multicellular fungi, amoebae, and oomycetes have been shown to function in various stress response (Messenguy and Dubois 2003; Ding et al. 2020; Rocha et al. 2016; Wang et al. 2018; Galardi-Castilla et al. 2013; Leesutthiphonchai and Judelson 2018). Thus, the regulation of stress responsive programs is possibly the ancestral function of MADS-box genes, which has been maintained in multicellular metazoans, both invertebrates and vertebrates (Vrailas-Mortimer et al. 2011; Blanchard et al. 2010; van der Linden et al. 2007; Potthoff and Olson 2007). The stress-responsive rather than housekeeping function of ancestral MADS-box genes could explain the observed gene loss in several extant lineages. Originated from an ancestral stress-responsive TF, SRF and MEF2 initially had presumably redundant functions upon duplication. Thus, the loss of SRF could have been compensated by MEF2 TFs in the unicellular ancestor of Archaeplastida. Supporting this assumption, the only MADS-box TF studied in microalgae, *Coccomyxa subellipsoidea* CsubMADS1, acts as a key regulator of stress tolerance (Nayar and Thangavel 2021). The colonization of the terrestrial habitat most likely required an expansion of the genetic toolkit regulating stress responses. Consistently, many MADS-box TFs have known function in regulating the response to stress, like FLOWERING LOCUS C (FLC), ARABIDOPSIS NITRATE REGULATED 1 (ANR1, AGL44) or AGL21 (Castelán-Muñoz et al. 2019).

In many unicellular organisms, the onset of reproduction is frequently induced by environmental stress, which may have facilitated the recruitment of MADS-box genes into the reproductive program (Piccirillo et al. 2015; Galardi-Castilla et al. 2013; Escalante et al. 2003; Leesutthiphonchai and Judelson 2018). In the land plant lineage, the evolution of spores and seeds that allow to withstand adverse environmental conditions may have been made possible by coupling the MADS regulation of resistance to reproductive

development. This is particularly evident in flowering plants, where MADS-box genes regulate floral patterning, but also the onset of flowering in responses to environmental cues (Castelán-Muñoz et al. 2019).

In summary, we conclude that the duplication of ancestral MEF2 type genes after plant landing gave rise to M-type and MIKC-type precursors that likely generated the genetic toolkit allowing both vegetative and reproductive programs sensing the environment, facilitating the evolution of complex structures adapted to the terrestrial environment. The varying sequences in the C-terminus of M-type and MIKC-type TFs largely increased the diversity of protein-protein interaction and thus the potential to form regulatory complexes. Multiple rounds of duplication and diversification of MIKC-type and M-type TFs likely have promoted the transition from a gametophyte-dominant to a sporophyte-dominant life cycle by equipping the sporophytic phase with developmental innovations such as flowers, fruits and seeds.

Interestingly, while animal MEF2 gene family did not expand as dramatically as the land plant orthologs, the increasing forms of splice variants have also largely increased the diversity of animal MEF2 TFs (Theissen et al. 1996). Animal MEF2 TFs are expressed predominantly within the early mesoderm (Potthoff and Olson 2007), which further differentiates into muscles and vascular and neuronal tissues, so that its function greatly promotes the mobility, integrity and sensibility of metazoans. Convergently, the plant MEF2 genes got recruited into body patterning and reproduction in response to environmental stimuli. Thus, during the evolution of multicellularity in both animals and plant, MEF2 type MADS-box TFs contributed to the formation of increasingly complex body plans adapted to the surrounding environment.

**Methods**

**Sequence and phylogenetic analyses**

To search for MADS-box proteins in the investigated species (Supplementary table 1), amino acid sequences of MADS-box proteins of *Arabidopsis thaliana*, human and yeast were used as queries in BLASTP programs. The output sequences were aligned to the MADS domain entries in the Conserved Domain Database (Lu et al. 2020) by the conserved domain search tool, CD-Search (Marchler-Bauer and Bryant 2004), which guided the extraction of MADS domains in each species. To inspect any missed MADS-box genes, HMM searches were carried out with HMMER (Eddy 2011). Genomes of interest were scanned against the MADS-box TF associated profile hidden Markov model (PF00319) retrieved from Pfam (now hosted by InterPro, http://www.ebi.ac.uk/interpro/) (Paysan-Lafosse et al. 2022).

MUSCLE was used to generate the amino acid alignments of MADS-box domains extracted from the selected sequences with default settings (Edgar 2004). We prepared two sets of alignments for the subsequent phylogenetic analyses: the alignments of only the conventional MADS-box region corresponding the first helix and the antiparallel strands; the alignments of the extended MADS-box domain definition to include the second helix additionally.

IQ-TREE 1.6.7 was applied to perform phylogenetic analyses for maximum likelihood trees (Nguyen et al. 2015). The implemented ModelFinder determined LG amino acid replacement matrix (Le and Gascuel 2008) to be the best substitution model in the tree inference (Kalyaanamoorthy et al. 2017). 1000 replicates of ultrafast bootstraps were applied to estimate the support for reconstructed branches (Hoang et al. 2018). MEGA11 was applied to generate neighbour-joining trees (Tamura et al. 2021), with p-distance (proportion of different amino acids), gamma distribution allowed for rate among sites and gaps treated by pairwise deletion. 1000 bootstrap replicates were generated and majority rule defines the consensus tree. Bayesian inference was carried out by Phylobayes (v3.2) under the CAT+GTR model with two chains. A consensus tree was built after the two chains were converged with the maxdiff less than 0.3 and the effective sample sizes of different parameters larger than 100 (Lartillot et al. 2009).

**Protein structure prediction and analyses**

We predicted the protein structures of selected MADS-box TFs by the web-based service ColabFold (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb) (Mirdita et al. 2022). The top ranked models were compared to resolved MADS-box protein structures, HsMEF2A (1EGW), HsSRF (1HBX) and AtSEP3 (7NB0), chains A respectively, downloaded from RCSB Protein Data Bank (https://www.rcsb.org/). The program "maxcluster" (http://www.sbg.bio.ic.ac.uk/~maxcluster/index.html) was used to perform structural comparisons based on computed TM-scores (Zhang and Skolnick 2005).

**Acknowledgements**

**Figures and legends**

**Fig.1** The models of MADS-box transcription factor (TF) evolution. **a.** The model by Theissen et al. (1996): Before the identification of M-type MADS-box TFs, known MADS-box TFs in plants were all belonging to the MIKC-type. The two types of MADS-box TFs in animals and fungi, SRF and MEF2, were considered duplicates that diverged after the split of the plant and animal lineages. **b.** The model by Alvarez-Buylla et al. (2000): Upon the identification of M-type MADS-box TFs in the genome of *Arabidopsis thaliana*, they were considered orthologous to SRF type TFs in animals and fungi, collectively named Type I MADS-box TFs, while MIKC-type in plants and MEF2 type in animals and fungi were clustered as Type II. Type I and II MADS-box genes originated by a hypothesized ancient duplication event predating the plant and animal divergence. **c.** The model proposed in

this study: With a broad survey across eukaryotes, the ancient duplication giving rise to SRF and MEF2 can be inferred as early as the origin of the most common recent ancestor of all living eukaryotic groups (consistent with model b). In several modern lineages both types have been retained; however, in the lineage of Archaeplastida, SRF type TFs got lost, and only MEF2 continued evolving. Subsequently, M-type and MIKC-type MADS-box TFs originated by gene duplication of a MEF2-type precursor in a common ancestor of all land plants. Black, primitive MADS-box TFs; yellow, SRF type; blue, MEF2 type; green, MIKC-type; cyan, M-type. Square, gene duplication event; cross, gene loss.

**Fig.2** Protein analyses of MADS-box TFs. **a.** Crystal structures (bold) of human SRF and MEF2A, and Arabidopsis SEP3 that were used as templates for structural comparisons (A-chains only), and AlphaFold2-predicted structural models of Archaeplastida MADS-box proteins including M-type and MIKC-type in land plants. Models were trimmed to the relevant structural segments matching the template structures. Structures were drawn using Pymol (Schrödinger and DeLano 2020). **b.c.** Similarity scores of all predicted models to human SRF and MEF2A, and Arabidopsis SEP3. SRF type (yellow) and MEF2 type (blue) were inferred by phylogeny. **d.** Similarity scores of predicted models for land plant MADS-box TFs to human SRF and MEF2A, and Arabidopsis AtSEP3.

**Fig.3** Maximum-likelihood (ML) tree of selected MADS-box domains (the first and second alpha helix and the connecting beta strands). Numbers for given branches of interest are support values: bootstrap values in ML trees (bold) / posterior probability in Bayesian inference (underlined) / bootstrap values in neighbor-joining trees. Branch color: yellow, SRF type; blue, MEF2 type; green, MIKC-type; cyan, M-type. Sequence IDs colored by category as listed on the right.

**Supplementary materials**

**Fig.S1 a.** Bayesian inference of phylogeny of surveyed MADS-box domains. Posterior probabilities are labelled next to branches of interest. **b.** Neighbor-joining tree of surveyed MADS-box domains. Bootstrap values are labelled next to branches of interest.

**Fig.S2** Similarity scores of predicted models in each taxonomic group to human SRF and MEF2, and Arabidopsis SEP3.

**Fig.S3** Maximum-Likelihood tree of surveyed MADS-box domains with SRF-like sequences in the core Chlorophytes (green clade pointed by the arrow). Bootstrap values are labelled above branches of interest.

**Fig.S4** Phylogenetic trees inferred with the alignment of the conventional definition of MADS domain (only the first helix and the beta strands). **a.** Maximum likelihood tree. Bootstrap values are labelled next to branches of interest. **b.** Bayesian inference of phylogeny. Posterior probabilities are labelled next to branches of interest. **c.** Neighbor-joining tree. Bootstrap values are labelled next to branches of interest.

**Table S1.** Species surveyed in this study.

**Table S2.** Protein similarity scores of AlphaFold2 predicted models for MADS-box TFs to human SRF and MEF2, and Arabidopsis AtSEP3.

**References**

Alvarez-Buylla ER., Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, Ribas de Pouplana L, Martínez-Castilla L, Yanofsky MF. (2000) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proceedings of the National Academy of Sciences of the United States of America 97(10):5328-5333. https://doi.org/10.1073/pnas.97.10.5328

Bemer M, Heijmans K, Airoldi C, Davies B, Angenent GC. (2010). An atlas of type I MADS box gene expression during female gametophyte and seed development in Arabidopsis. Plant physiology 154(1):287-300. https://doi.org/10.1104/pp.110.160770

Blanchard FJ, Collins B, Cyran SA, Hancock DH, Taylor MV, Blau J. (2010) The transcription factor Mef2 is required for normal circadian behavior in Drosophila. The Journal of neuroscience 30(17):5855-5865. https://doi.org/10.1523/JNEUROSCI.2688-09.2010

Burki F, Roger AJ, Brown MW, Simpson A. (2020) The New Tree of Eukaryotes. Trends in Ecology and Evolution 35(1):43-55. https://doi.org/10.1016/j.tree.2019.08.008

Castelán-Muñoz N, Herrera J, Cajero-Sánchez W, Arrizubieta M, Trejo C, García-Ponce B, Sánchez MP, Álvarez-Buylla ER, Garay-Arroyo A. (2019) MADS-Box Genes Are Key Components of Genetic Regulatory Networks Involved in Abiotic Stress and Plastic Developmental Responses in Plants. Frontiers in Plant Science 10:853. https://doi.org/10.3389/fpls.2019.00853

De Bodt S, Raes J, Van de Peer Y, Theissen G. (2003). And then there were many: MADS goes genomic. Trends in Plant Science 8(10):475-483. https://doi.org/10.1016/j.tplants.2003.09.006

Ding Z, Xu T, Zhu W, Li L, Fu Q. (2020) A MADS-box transcription factor FoRlm1 regulates aerial hyphal growth, oxidative stress, cell wall biosynthesis and virulence in Fusarium oxysporum f. sp. cubense. Fungal Biology 124(3-4):183-193. https://doi.org/10.1016/j.funbio.2020.02.001

Eddy SR. (2011) Accelerated profile HMM searches. PLoS Computational Biology 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar RC. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5):1792-1797. https://doi.org/10.1093/nar/gkh340

Escalante R, Moreno N, Sastre L. (2003) Dictyostelium discoideum developmentally regulated genes whose expression is dependent on MADS box transcription factor SrfA. Eukaryotic Cell 2(6):1327-1335. https://doi.org/10.1128/EC.2.6.1327-1335.2003

Galardi-Castilla M, Fernandez-Aguado I, Suarez T, Sastre L. (2013) Mef2A, a homologue of animal Mef2 transcription factors, regulates cell differentiation in Dictyostelium discoideum. BMC developmental biology 13:12. https://doi.org/10.1186/1471-213X-13-12

Gramzow L, Ritz MS, Theissen G. (2010) On the origin of MADS-domain transcription factors. Trends in Genetics 26(4):149-153. https://doi.org/10.1016/j.tig.2010.01.004

Gramzow L, Theissen G. (2013) Phylogenomics of MADS-Box genes in plants—two opposing life styles in one gene family. Biology (Basel) 2(3):1150–1164. https://doi.org/10.3390/biology2031150

Gramzow L, Weilandt L, Theißen G. (2014) MADS goes genomic in conifers: towards determining the ancestral set of MADS-box genes in seed plants. Annals of Botany 114(7):1407-1429. https://doi.org/10.1093/aob/mcu066

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. (2018) UFBoot2: improving the ultrafast bootstrap approximation. Molecular Biology and Evolution 35(2): 518-522. https://doi.org/10.1093/molbev/msx281

Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, Seo M, Sato S, Yamada T, Mori H, Tajima N, Moriyama T, Ikeuchi M, Watanabe M, Wada H, Kobayashi K, Saito M, Masuda T, Sasaki-Sekimoto Y, Mashiguchi K, Awai K, Shimojima M, Masuda S, Iwai M, Nobusawa T, Narise T, Kondo S, Saito H, Sato R, Murakawa M, Ihara Y, Oshima-Yamada Y, Ohtaka K, Satoh M, Sonobe K, Ishii M, Ohtani R, Kanamori-Sato M, Honoki R, Miyazaki D, Mochizuki H, Umetsu J, Higashi K, Shibata D, Kamiya Y, Sato N, Nakamura Y, Tabata S, Ida S, Kurokawa K, Ohta H. (2014) Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. Nature Communications 5:3978. https://doi.org/10.1038/ncomms4978

Kalyaanamoorthy S, Minh BQ, Wong T, von Haeseler A, Jermiin LS. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14(6):587-589. https://doi.org/10.1038/nmeth.4285

Kaufmann, K., Melzer, R., and Theissen, G. (2005) MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. Gene 347:183-198. https://doi.org/10.1016/j.gene.2004.12.014

Kofuji R, Sumikawa N, Yamasaki M, Kondo K, Ueda K, Ito M, Hasebe M. (2003) Evolution and divergence of the MADS-box gene family based on genome-wide expression analyses. Molecular Biology and Evolution, 20(12):1963-1977. https://doi.org/10.1093/molbev/msg216

Lai X, Vega-Léon R, Hugouvieux V, Blanc-Mathieu R, van der Wal F, Lucas J, Silva CS, Jourdain A, Muino JM, Nanao MH, Immink R, Kaufmann K, Parcy F, Smaczniak C, Zubieta C. (2021) The intervening domain is required for DNA-binding and functional identity of plant MADS transcription factors. Nature Communications 12(1):4760. https://doi.org/10.1038/s41467-021-24978-w

Lartillot N, Lepage T, Blanquart S. (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25(17):2286-2288. https://doi.org/10.1093/bioinformatics/btp368

Le SQ, Gascuel O. (2008) An improved general amino acid replacement matrix. Molecular Biology and Evolution 25(7):1307-1320. https://doi.org/10.1093/molbev/msn067

Leesutthiphonchai W, Judelson HS. (2018) A MADS-box transcription factor regulates a central step in sporulation of the oomycete Phytophthora infestans. Molecular Microbiology 110(4):562-575. https://doi.org/10.1111/mmi.14114

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. (2020) CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Research 48(D1):D265-D268. https://doi.org/10.1093/nar/gkz991

Marchler-Bauer A, Bryant SH. (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Research 32:W327-W331. https://doi.org/10.1093/nar/gkh454

Masiero S, Colombo L, Grini PE, Schnittger A, Kater MM. (2011). The emerging importance of type I MADS box transcription factors for plant reproduction. The Plant cell 23(3):865-872. https://doi.org/10.1105/tpc.110.081737

Messenguy F, Dubois, E. (2003) Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. Gene 316:1-21. https://doi.org/10.1016/s0378-1119(03)00747-9

Nam J, dePamphilis CW, Ma H, Nei M. (2003) Antiquity and evolution of the MADS-box gene family controlling flower development in plants. Molecular Biology and Evolution 20(9):1435-1447. https://doi.org/10.1093/molbev/msg152

Nam J, Kim J, Lee S, An G, Ma H, Nei M. (2004) Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proceedings of the National Academy of Sciences of the United States of America 101(7): 1910-1915. https://doi.org/10.1073/pnas.0308430100

Nayar S, Thangavel G. (2021) CsubMADS1, a lag phase transcription factor, controls development of polar eukaryotic microalga Coccomyxa subellipsoidea C-169. The Plant journal 107(4):1228-1242. https://doi.org/10.1111/tpj.15380

Ng M, Yanofsky MF. (2001). Function and evolution of the plant MADS-box gene family. Nature Reviews Genetics 2(3):186-195. https://doi.org/10.1038/35056041

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(1):268-274. https://doi.org/10.1093/molbev/msu300

Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas FB, Vanderstraeten L, Becker D, Lang D, Vosolsobě S, Rombauts S, Wilhelmsson PKI, Janitza P, Kern R, Heyl A, Rümpler F, Villalobos LIAC, Clay JM, Skokan R, Toyoda A, Suzuki Y, Kagoshima H, Schijlen E, Tajeshwar N, Catarino B, Hetherington AJ, Saltykova A, Bonnot C, Breuninger H, Symeonidi A, Radhakrishnan GV, Van Nieuwerburgh F, Deforce D, Chang C, Karol KG, Hedrich R, Ulvskov P, Glöckner G, Delwiche CF, Petrášek J, Van de Peer Y,

Friml J, Beilby M, Dolan L, Kohara Y, Sugano S, Fujiyama A, Delaux PM, Quint M, Theißen G, Hagemann M, Harholt J, Dunand C, Zachgo S, Langdale J, Maumus F, Van Der Straeten D, Gould SB, Rensing SA. (2018) The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. Cell 174(2):448-464.e24. https://doi.org/10.1016/j.cell.2018.06.033

Parenicová L, de Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, Angenent GC, Colombo L. (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. The Plant Cell 15(7):1538-1551. https://doi.org/10.1105/tpc.011544

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunić I, Marchler-Bauer A, Mi H, Natale DA, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A. (2022) InterPro in 2022. Nucleic Acids Research gkac993. Advance online publication. https://doi.org/10.1093/nar/gkac993

Piccirillo S, Morales R, White MG, Smith K, Kapros T, Honigberg SM. (2015) Cell Differentiation and Spatial Organization in Yeast Colonies: Role of Cell-Wall Integrity Pathway. Genetics 201(4):1427-1438. https://doi.org/10.1534/genetics.115.180919

Potthoff M, Olson EN. (2007) MEF2: a central regulator of diverse developmental programs. Development 134(23):4131-4140. doi: https://doi.org/10.1242/dev.008367

Qiu Y, Köhler C. (2022) Endosperm Evolution by Duplicated and Neofunctionalized Type I MADS-Box Transcription Factors. Molecular Biology and Evolution 39(1):msab355. https://doi.org/10.1093/molbev/msab355

Rocha MC, Fabri JH, Franco de Godoy K, Alves de Castro P, Hori JI, Ferreira da Cunha A, Arentshorst M, Ram AF, van den Hondel CA, Goldman GH, Malavazi I. (2016) Aspergillus fumigatus MADS-Box Transcription Factor rlmA Is Required for Regulation of the Cell Wall

Integrity and Virulence. G3 (Bethesda, Md.) 6(9):2983-3002. https://doi.org/10.1534/g3.116.031112

Schrödinger L, DeLano W. (2020) PyMOL. Retrieved from http://www.pymol.org/pymol

Schwarz-Sommer Z, Huijser P, Nacken W, Saedler H, Sommer H. (1990) Genetic control of flower development by homeotic genes in Antirrhinum majus. Science 250:931-936. https://www.science.org/doi/10.1126/science.250.4983.931

Shore P, Sharrocks AD. (1995) The MADS-Box Family of Transcription Factors. European Journal of Biochemistry 229:1-13. https://doi.org/10.1111/j.1432-1033.1995.0001l.x

Tamura K, Stecher G, Kumar S. (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution 38(7):3022-3027. https://doi.org/10.1093/molbev/msab120

Thangavel G, and Nayar S. (2018) A Survey of MIKC Type MADS-Box Genes in Non-seed Plants: Algae, Bryophytes, Lycophytes and Ferns. Frontiers in Plant Science 9:510. https://doi.org/10.3389/fpls.2018.00510

Theissen G, Kim JT, Saedler, H. (1996) Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. Journal of molecular evolution 43(5):484-516. https://doi.org/10.1007/BF02337521

van der Linden AM, Nolan KM, Sengupta P. (2007) KIN-29 SIK regulates chemoreceptor gene expression via an MEF2 transcription factor and a class II HDAC. The EMBO journal 26(2):358-370. https://doi.org/10.1038/sj.emboj.7601479
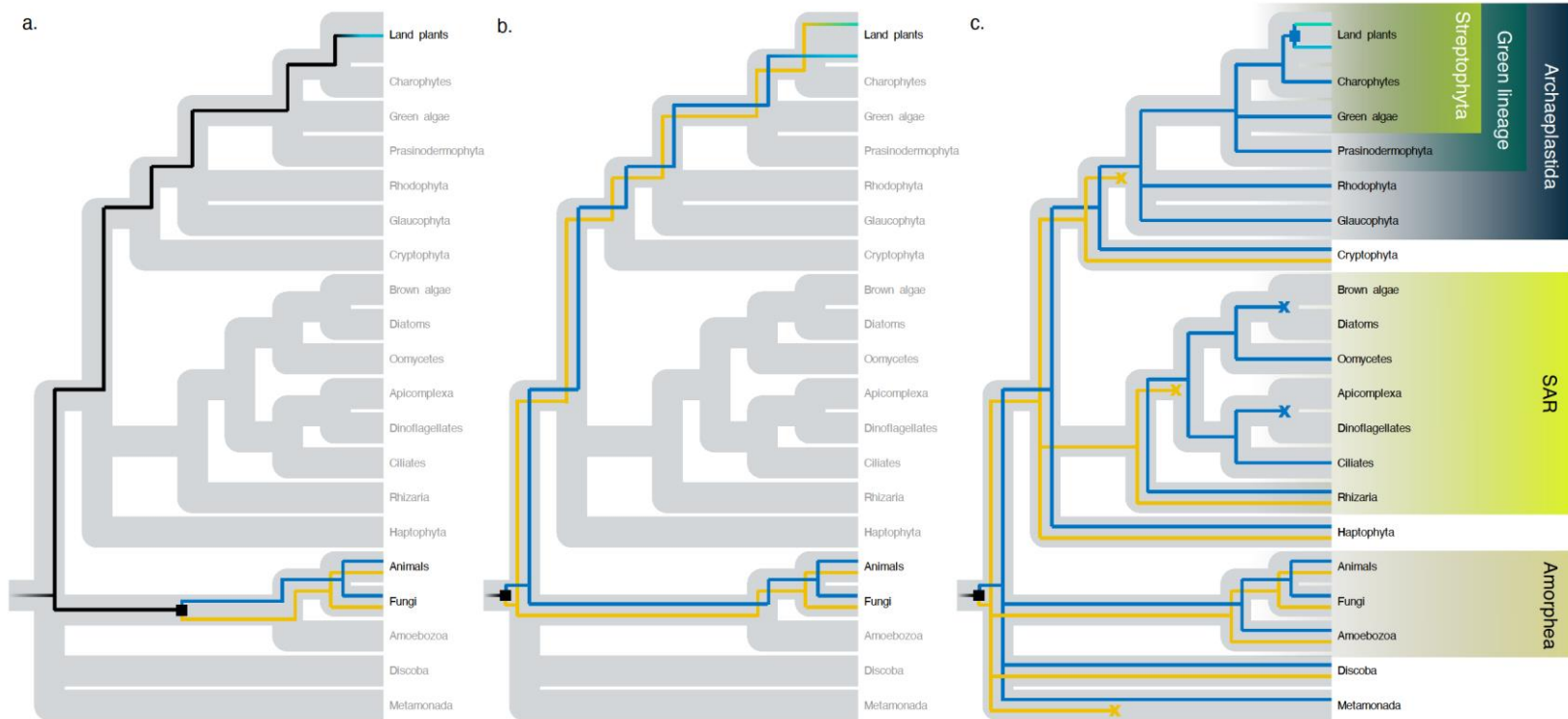
Vrailas-Mortimer A, del Rivero T, Mukherjee S, Nag S, Gaitanidis A, Kadas D, Consoulas C, Duttaroy A, Sanyal S. (2011) A muscle-specific p38 MAPK/Mef2/MnSOD pathway regulates stress, motor function, and life span in Drosophila. Developmental Cell 21(4):783-795. https://doi.org/10.1016/j.devcel.2011.09.002
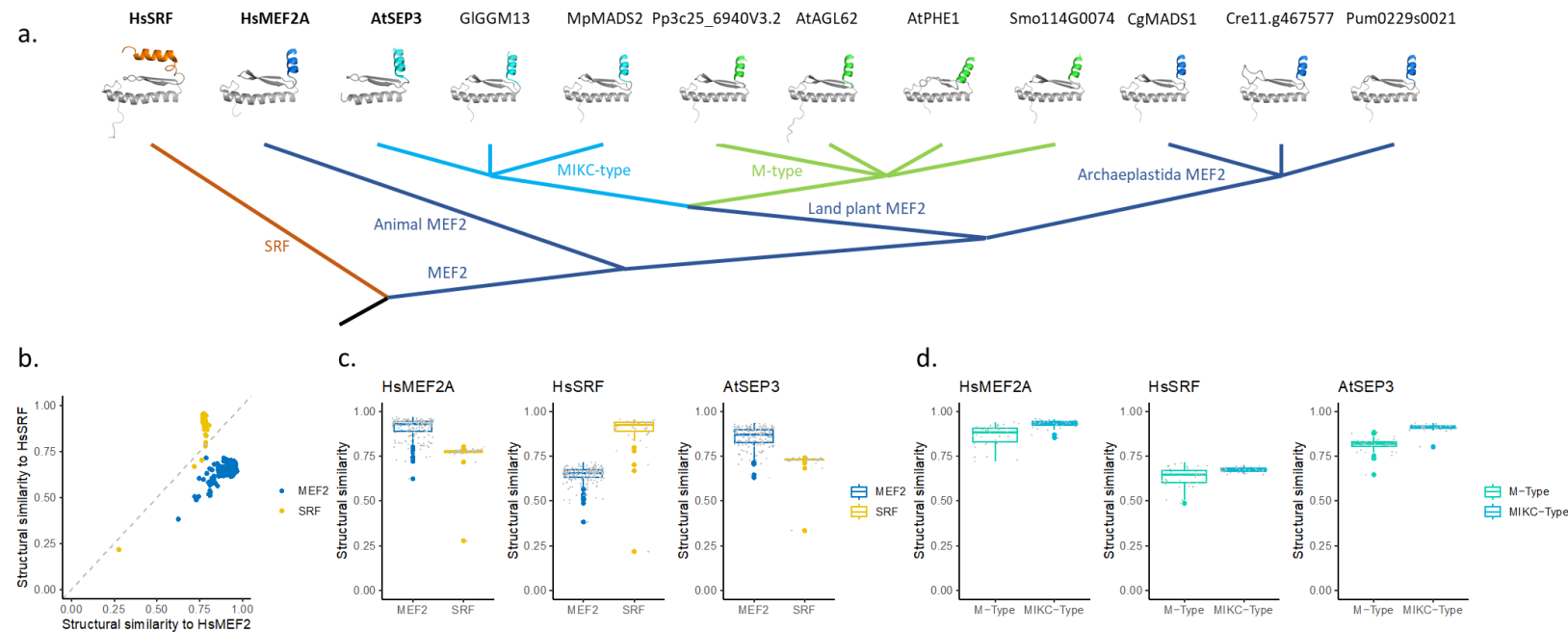
Wang Q, Du M, Wang S, Liu L, Xiao L, Wang L, Li T, Zhuang H, Yang E. (2018) MADS-Box Transcription Factor MadsA Regulates Dimorphic Transition, Conidiation, and Germination of Talaromyces marneffei. Frontiers in Microbiology 9:1781. https://doi.org/10.3389/fmicb.2018.01781

Zhang Y, Skolnick J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33:2302-2309.

**Fig.1** The models of MADS-box transcription factor (TF) evolution. **a.** The model by Theissen et al. (1996): Before the identification of M-type MADS-box TFs, known MADS-box TFs in plants were all belonging to the MIKC-type. The two types of MADS-box TFs in animals and fungi, SRF and MEF2, were considered duplicates that diverged after the split of the plant and animal lineages. **b.** The model by Alvarez-Buylla et al. (2000): Upon the identification of M-type MADS-box TFs in the genome of *Arabidopsis thaliana*, they were considered orthologous to SRF type TFs in animals and fungi, collectively named Type I MADS-box TFs, while MIKC-type in plants and MEF2 type in animals and fungi were clustered as Type II. Type I and II MADS-box genes originated by a hypothesized ancient duplication event predating the plant and animal divergence. **c.** The model proposed in this study: With a broad survey across eukaryotes, the ancient duplication giving rise to SRF and MEF2 can be inferred as early as the origin of the most common recent ancestor of all living eukaryotic groups (consistent with model b). In several modern lineages both types have been retained; however, in the lineage of Archaeplastida, SRF type TFs got lost, and only MEF2 continued evolving. Subsequently, M-type and MIKC-type MADS-box TFs originated by gene duplication of a MEF2-type precursor in a common ancestor of all land plants. Black, primitive MADS-box TFs; yellow, SRF type; blue, MEF2 type; green, MIKC-type; cyan, M-type. Square, gene duplication event; cross, gene loss.

**Fig.2** Protein analyses of MADS-box TFs. **a.** Crystal structures (bold) of human SRF and MEF2A, and Arabidopsis SEP3 that were used as templates for structural comparisons (A-chains only), and AlphaFold2-predicted structural models of Archaeplastida MADS-box proteins including M-type and MIKC-type in land plants. Models were trimmed to the relevant structural segments matching the template structures. Structures were drawn using Pymol (Schrödinger and DeLano 2020). **b.c.** Similarity scores of all predicted models to human SRF and MEF2A, and Arabidopsis SEP3. SRF type (yellow) and MEF2 type (blue) were inferred by phylogeny. **d.** Similarity scores of predicted models for land plant MADS-box TFs to human SRF and MEF2A, and Arabidopsis AtSEP3.

**Fig.3** Maximum-likelihood (ML) tree of selected MADS-box domains (the first and second alpha helix and the connecting beta strands). Numbers for given branches of interest are support values: bootstrap values in ML trees (bold) / posterior probability in Bayesian inference (underlined) / bootstrap values in neighbor-joining trees. Branch color: yellow, SRF type; blue, MEF2 type; green, MIKC-type; cyan, M-type. Sequence IDs colored by category as listed on the right.