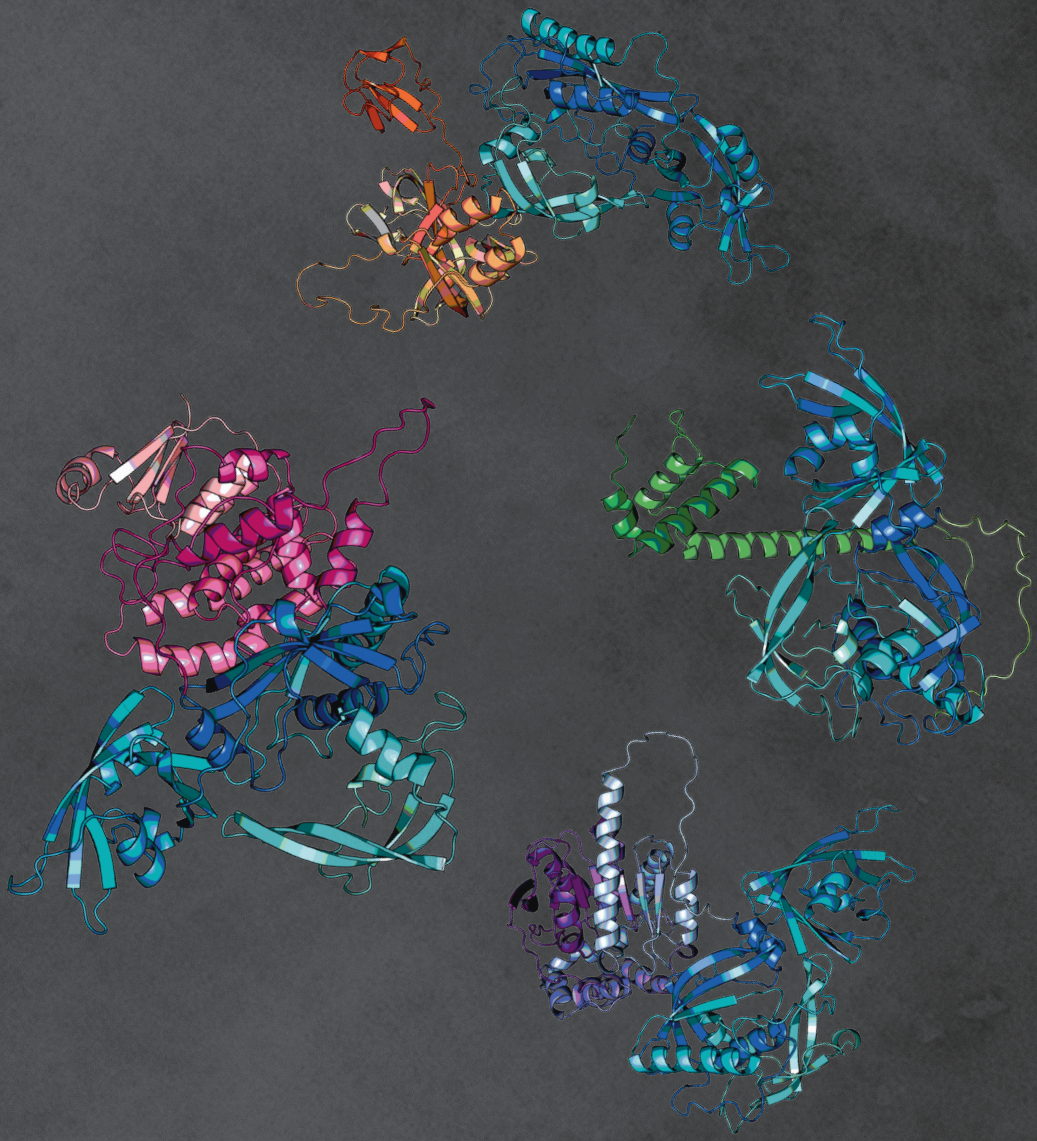
CONSERVATION OF ARF DNA-BINDING DURING EVOLUTION



Juriaan Rienstra

PROPOSITIONS

1. Functional constraints on ARF DNA-binding resulted in conserved DNA-binding preferences over 800 million years of streptophyte evolution.
   (this thesis)

2. Protein sequence comparisons are meaningless without consideration of the encoded protein structure.
   (this thesis)

3. The dramatic increase of high quality predicted protein structures led to an equally impressive increase in the number of structural biology “experts”.

4. RuBisCO is not inefficient at carbon fixation, it operates as efficiently as it can within the confines of Muller’s ratchet.

5. Large language models like ChatGPT help democratize science for non-native English speakers.

6. Transitioning from voluntary philantropy to mandatory wealth taxes is necessary to democratically define “good causes”.

7. It is more beneficial to turn politicians (or policy makers) into scientists than scientists into politicians (or policy makers).

Propositions belonging to the thesis, entitled

Conservation of ARF DNA-binding during evolution

Juriaan Rienstra

Wageningen, 30 April 2024

Conservation of ARF DNA-binding during evolution

Juriaan Rienstra

Thesis committee

Promotor

Prof. Dr Dolf Weijers

Professor of Biochemistry

Wageningen University & Research

Co-promotor

Dr Jorge Hernández-García

Post-doctoral researcher at the Laboratory of Biochemistry

Wageningen University & Research

Other members

Prof. Dr Richard Immink, Wageningen University & Research

Prof. Dr Berend Snel, Utrecht University

Prof. Dr Kerstin Kaufmann, Humboldt University, Germany

Dr Chloe Zubieta, CEA Grenoble, France

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences

Conservation of ARF DNA-binding during evolution

Juriaan Rienstra

Thesis

submitted in fulfilment of the requirement for the degree of doctor

at Wageningen University

By the authority of the Rector Magnificus

Prof. Dr C. Kroeze,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Tuesday 30 April 2024

at 1:30 pm in Omnia.

Juriaan Rienstra

Conservation of ARF DNA-binding during evolution

168 Pages

PhD Thesis, Wageningen University & Research, Wageningen, the Netherlands (2024)

With references, with summary in English

ISBN: 978-94-6469-891-6

DOI: 10.18174/647747

Table of contents

Chapter 1

General Introduction **7**

Chapter 2

To bind or not to bind: how Auxin Response Factors select their target genes **17**

Chapter 3

Conservation and diversity in sequence and structure of the ARF transcription factors **34**

Chapter 4

DNA-binding preferences of the *Marchantia polymorpha* ARF1 transcription factor **57**

Chapter 5

Evolvability of the ARF-DNA interface **79**

Chapter 6

ARF-B3 domain DNA-binding conservation across ARF classes and streptophyte lineages **115**

Chapter 7

General Discussion **135**

References **145**

Summary **155**

Acknowledgements **157**

Curriculum Vitae **162**

Publications **163**

Education statement **165**

Chapter

1 General Introduction

The evolution of TF DNA-binding mechanisms

The growth and development of any organism require responding to internal and external signals by switching pathways and processes on or off. The proteins responsible for these switches are transcription factors (TFs), proteins that can bind DNA and activate or repress downstream genes by recognizing distinct DNA elements. Because of their importance, the mechanisms by which they bind DNA have been actively studied. TFs evolve complex, three-dimensional interaction mechanisms to recognize their cognate DNA elements and indicate there is no simple 'code' that all TFs follow to bind DNA. However, from the TF crystal structures two general mechanisms have been derived, both of which are utilized by TFs to varying degrees: base- and shape-readout, as reviewed by (Slattery *et al.*, 2014)**(Fig. 1)**. Base readout refers to the physical interactions between the TF residues and the nucleobases, often involving the formation of hydrogen bonds **(Fig. 1A)** (Luscombe *et al.*, 2001; Slattery *et al.*, 2014). Shape readout refers to recognizing intrinsic properties of the double stranded DNA **(Fig. 1B)** (Abe *et al.*, 2015; Andrews *et al.*, 2023; Slattery *et al.*, 2014; Zeiske *et al.*, 2018). The strongly conserved TATA-box Binding Proteins (TBP, **Fig. 1B**) are an example of TFs that almost exclusively rely on shape readout since they recognize AT-rich DNA-elements by the ease of bending of the DNA, rather than via base readout (Patikoglou *et al.*, 1999). Other TFs generally recognize their cognate DNA sites via a combination of base and shape readout. Besides these intrinsic recognition mechanisms, TF specificity can be influenced via post-translational modifications, for example, acetylation of DNA-binding residues (Louphrasitthiphol *et al.*, 2020), or by DNA modifications, such as cytosine methylation (Hashimoto *et al.*, 2016; Yin *et al.*, 2017). Eukaryotic TFs generally do not bind by themselves but can form homodimers to cooperatively bind mirrored bipartite sites (Fontana *et al.*, 2023; Hames *et al.*, 2008; Sloan *et al.*, 2020), form intrafamilial heterodimers to bind composite motifs (Aerts *et al.*, 2018; Chen *et al.*, 2019; Lai *et al.*, 2020) and DNA-specificity can be influenced by interactions with co-factors (Lai *et al.*, 2020; Rieu *et al.*, 2023; Slattery *et al.*, 2011). Historically, biochemical methods have been designed to derive the highest affinity elements of TFs. However, it has become clear that low and medium-affinity sites play major roles in development (Crocker *et al.*, 2015; Kribelbauer *et al.*, 2019). Indeed, mathematical models suggest that for a given TF, slight promiscuity towards low and medium affinity sites increases the evolutionary chance of gaining new target genes, which could offer evolutionary advantages in both adaptation and speciation (Berg *et al.*, 2004; Tugrul *et al.*, 2015). Together, although DNA-binding of any specific TF can be explained by two general biochemical mechanisms, understanding how DNA-binding is affected *in vivo* requires investigations of multiple angles, including evolution of the TF.

The evolution of TFs with new DNA-binding specificity often follows gene or whole genome duplications, and the novel TFs are often associated with the appearance of novel tissues and cell types (Jolma *et al.*, 2013; Nitta *et al.*, 2015). If TFs regulate

critical processes, such as the conserved *Hox* TFs that are central to embryogenesis in both mammals and insects (Wanninger, 2024), they experience strong evolutionary constraints to retain the same DNA-binding specificity across evolution. These constraints likely occur on the majority of TFs given that, apart from the Cys2 His2 Zinc Finger (C2H2-ZF) family (Lambert *et al.*, 2019), the vast majority of orthologous TFs maintain similar DNA-binding specificities in humans and fruit flies (*Drosophila*) after 600 million years of evolution (Gera *et al.*, 2022; Jolma *et al.*, 2013; Nitta *et al.*, 2015; Reece-Hoyes *et al.*, 2013). Therefore, the evolutionary trigger for generating novel DNA-binding specificities are gene or whole genome duplication events that cause the duplication of an ancestral gene into two paralogous copies (Schmitz *et al.*, 2016). These copies may then undergo one of two evolutionary processes. The first process is subfunctionalization, wherein the ancestral function is split among the copies and the paralogs become specialized. This could occur via changes in expression profiles, interactions partners, or coevolution of paralogs with their set of target genes (Gera *et al.*, 2022; Perez *et al.*, 2014; Schmitz *et al.*, 2016). The second process is neofunctionalization when the ancestral function is maintained in one paralog and the other paralog acquires a new role, such as novel DNA-binding specificity (Gera *et al.*, 2022; Schmitz *et al.*, 2016). These TFs with novel DNA-

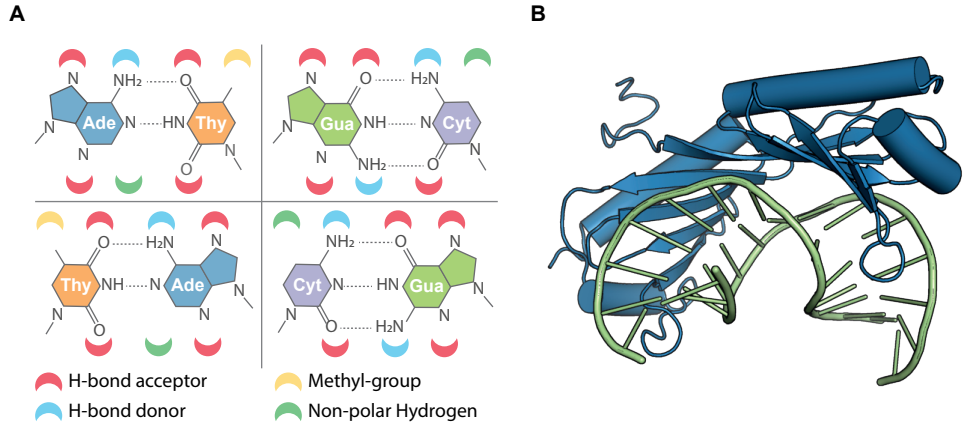


Fig. 1 | Base and Shape readout mechanisms. (A) Base readout mechanisms describe direct interactions between the chemical groups of the nucleobases (moon-shape) and the transcription factor (TF) residues (not shown). Whereas the pattern of hydrogen bond (H-bond) acceptor (red), H-bond donors (blue), hydrophobic methyl group (yellow) and heterocyclic non-polar hydrogen (green) in the major groove (top rows, four chemical groups) are distinct, the chemical groups in the minor groove are not (lower rows, three chemical groups). Image based on (Slattery *et al.*, 2014). **(B)** Shape readout mechanisms describe interactions between the TF and the DNA that rely on global and local DNA features such as conformational flexibility. TATA-box Binding Proteins (TBP, blue) bind A:T and T:A rich sequences in the minor groove and cause large conformational changes in the DNA (green), including bending and widening of the minor groove. As TBPs make very little non-covalent bonds in the minor groove, their DNA-specificity comes from the ease at which a DNA site adopts the extreme conformation. PDB: 1qna.

binding can therefore evolve to regulate a unique set of target genes and are therefore associated with the appearance of new tissues and cell types, such as the mammalian-specific TFs ATF2 and E2F that are important for placental morphogenesis (Nitta *et al.*, 2015). TF neofunctionalization also occurs for plant TFs, such as the MYB TF *DUO POLLEN1* (*DUO1*), which gained novel DNA-binding – as compared with its closest relative *MYB21* – and targets a set of conserved genes responsible for sperm-cell differentiation in extant land plants (Higo *et al.*, 2016). Although gaining novel DNA-binding specificities are one way for neofunctionalizing TFs to acquire new functions, it should be mentioned that sub- and neofunctionalization of TFs can also occur without changing DNA-binding, for example by changes in protein-protein interactions, changes in expression profiles, protein domain rearrangements, and only partial loss or acquisition of target genes (Reece-Hoyes *et al.*, 2013; Schmitz *et al.*, 2016; Tu *et al.*, 2022; Zhang *et al.*, 2018). Although all TF families will therefore experience the same evolutionary forces, how these forces drive changes should be studied on the level of the individual TF family.

Engineering and artificial evolution of TF DNA-binding specificity

Given the central role of TFs in switching pathways on and off, they offer tremendous potential in synthetic biology, pharmacology and agronomy (Deng *et al.*, 2022). However, since most TFs have evolved very intricate and specific atomic mechanisms to bind DNA, rational design for altered or novel DNA-binding has proven difficult. For example, for Cpf1, a prokaryotic RNA-guided endonuclease, mutants with altered DNA-binding specificities were generated via error-prone PCR, however, their protein structures revealed subtle structural rearrangements that were not readily predictable from the wild-type structure (Gao *et al.*, 2017; Nishimasu *et al.*, 2017). Rational design of TF binding is therefore difficult, with two exceptions: the prokaryotic Transcription Activator-Like Effectors (TALE) and the C2H2-ZF domains. These domains can recognize one (TALE) or three (C2H2-ZF) basepairs max, but since TALE and C2H2-ZF proteins are often tandem repeats of these domain, they could target any possible DNA sequence via domain rearrangements (Mueller *et al.*, 2021; Persikov *et al.*, 2015; Sanjana *et al.*, 2012). For these reasons, both families have been extensively used in synthetic biology (Chandrasegaran and Carroll, 2015; Osakabe and Osakabe, 2015). Notably, TALE proteins are the only identified DNA-binding proteins that conform to the 'code' hypothesis of DNA recognition (Slattery *et al.*, 2014). Successful efforts in engineering TFs came from directed evolution experiments, which was used to change the T7 RNA Polymerase (RNAP) DNA-binding specificity towards a novel DNA element (Dickinson *et al.*, 2013; Esvelt *et al.*, 2011). By using two different intermediate DNA elements, they could simulate different evolutionary 'tracks' and showed that the engineered T7 RNAP from the first track ended with much lower transcriptional activity than the T7 RNAP from the second track, i.e.

the first track resulted in an evolutionary ratchet. The evolutionary path can therefore influence how TFs acquire novel DNA-binding, and cannot be easily predicted *de novo*. For extant TFs however, the evolutionary path can be traced successfully by ancestral reconstruction (Carroll *et al.*, 2011; Starr *et al.*, 2017). Together, subtle structural effects and the unpredictable nature of evolution make rational design of novel TFs difficult, but these attempts nevertheless reveal much about protein structure and evolution.

Auxin and the origin of land plants

The plant hormone auxin, indole-3-acetic acid, has been tied with major morphological changes during land plant evolution and is involved in virtually all land plant growth and developmental processes (Bowman *et al.*, 2021; Li *et al.*, 2016; Weijers and Wagner, 2016). At the level of individual cells, auxin signalling is transduced by the nuclear auxin pathway (NAP), which is composed of three protein families conserved across land plants: the auxin receptors TRANSPORT INHIBITOR 1/AUXIN SIGNALING F-BOX (TIR1/AFB), the transcriptional co-repressors AUXIN/INDOLE-3-ACETIC ACID (Aux/IAA) and the transcription factors (TFs) AUXIN RESPONSE FACTORs (ARFs). TIR1/AFB proteins are E3 ubiquitin ligases and when auxin binds to the TIR1/AFB proteins, it promotes interactions with the Aux/IAA proteins, leading to the ubiquitination of Aux/IAA and their subsequent proteolysis by the 26S proteasome (Suzuki *et al.*, 2023; Tan *et al.*, 2007). Aux/IAA proteins act as transcriptional repressors via the recruitment of the repressor protein TOPLESS (TPL). Aux/IAA do not bind DNA directly but interact with ARFs and thereby impose transcriptional repression (Kato *et al.*, 2015; Lavy *et al.*, 2016; Tiwari, 2001). Upon TIR1/AFB-mediated proteolysis of the Aux/IAA proteins, the ARFs are liberated to regulate gene expression. In land plants, the ARFs can be divided into three subfamilies: A, B and C-ARFs of which the A-ARFs are transcriptional activators and the B-ARFs transcriptional repressors (Kato *et al.*, 2015; Tiwari *et al.*, 2003). The activity of the C-ARFs is unclear, as discussed by (Kato *et al.*, 2018). ARFs bind to auxin response elements (AuxRE) for which angiosperm (Arabidopsis and maize) A and B-ARFs strongly prefer AuxRE composed of TGTCGG (Boer *et al.*, 2014; Galli *et al.*, 2018; Stigliani *et al.*, 2019). The interactions between auxin, TIR1/AFB, Aux/IAA and the transcriptional activity of the ARFs thereby represent a dynamic signalling pathway dependent on auxin. However, regardless of these dynamics, the downstream genes that become auxin-responsive are ultimately decided by the ARFs and their AuxRE-binding preferences. The ARFs are therefore key to plant growth and development, and understanding their DNA-binding preferences and the evolution thereof are crucial to understanding how the NAP evolved to become involved in almost all developmental processes in land plants.

Phylogenetic studies have revealed the origin and evolution of the NAP across the streptophyta – covering land plants and their closest algal relatives **(Fig. 2)** (Bowman *et al.*, 2021; Carrillo-Carrasco *et al.*, 2023; Flores-Sandoval *et al.*, 2018a; Kato *et al.*, 2018;

Martin-Arevalillo *et al.*, 2019; Mutte *et al.*, 2018). The ancestor to all streptophytes likely had a single proto-ARF with extant C-ARF characteristics in DNA-binding preferences (Flores-Sandoval *et al.*, 2018a; Martin-Arevalillo *et al.*, 2019). This proto-ARF duplicated and formed the A/B-ARF and C-ARF clades, both clades which are found in extant streptophyte algae (Carrillo-Carrasco *et al.*, 2023; Mutte *et al.*, 2018). Although these algae encode TIR1/AFB-like and Aux/IAA-like proteins, both lack the residues critical for interacting with auxin (Bowman *et al.*, 2017; Mutte *et al.*, 2018). The last common ancestor (LCA) of the land plants (embryophytes) first established a fully functional NAP, which included the duplication of the A/B-ARF into A and B-ARFs (Bowman *et al.*, 2021; Flores-Sandoval *et al.*, 2018a; Kato *et al.*, 2018; Mutte *et al.*, 2018). The acquisition of auxin as a hormone and the evolution of a functional signalling pathway were therefore central to terrestrialization by land plants. The liverwort *Marchantia polymorpha* (bryophytes), never underwent full genome duplications and with a single copy of TIR1/AFB (Mp*TIR1*), Aux/IAA (Mp*IAA*) and one A, B and C-ARF (Mp*ARF1*, Mp*ARF2* and Mp*ARF3* respectively) represents the ancestral state of the NAP (Bowman *et al.*, 2017; Flores-Sandoval *et al.*, 2015; Kato *et al.*, 2015; Kato *et al.*, 2018).

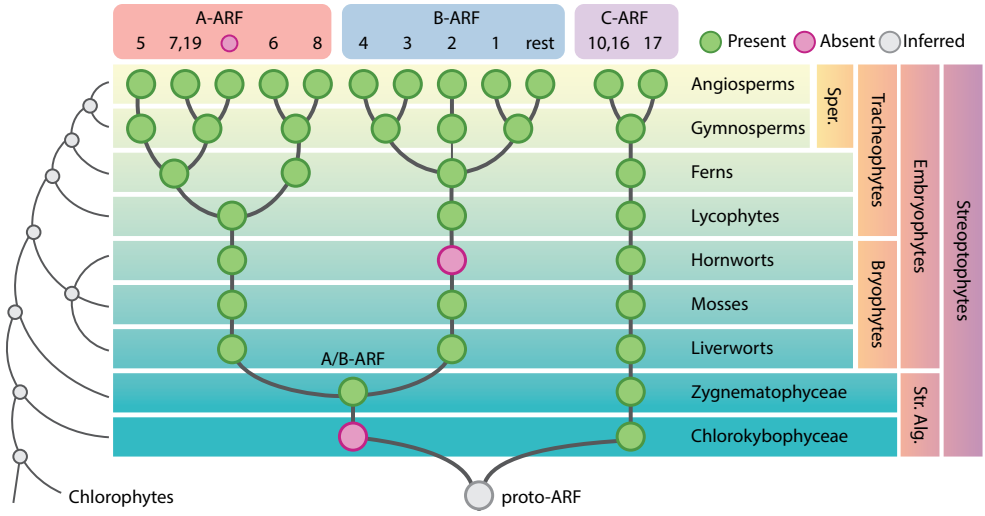


Fig. 2 | Evolution of the AUXIN RESPONSE FACTOR (ARF) family across streptophytes. Left: schematic phylogenetic tree of taxonomic classes, based on (McCourt *et al.*, 2023). Middle: schematic phylogenetic tree of ARF family in streptophytes, based on (Carrillo-Carrasco *et al.*, 2023; Mutte *et al*., 2018). Hornworts independently lost B-ARFs. Top: Numbers indicate which genes of Arabidopsis thaliana (Arabidopsis) belong to each subfamily and magenta circle indicates a missing subfamily in Arabidopsis. Right: indicates major taxonomic classifications. Abbreviations: streptophyte algae (Str. Alg.), spermatophytes (Sper., seed plants). Note that several lineages of streptophyte algae were omitted for brevity, see (Carrillo-Carrasco *et al*., 2023; McCourt *et al.*, 2004, 2023).

In contrast, the NAP has been primarily studied in flowering plants (angiosperms), chief among them *Arabidopsis thaliana* (Arabidopsis), in which whole genome duplications led to a large expansion of NAP components. With 5 TIR1/AFB, 28 Aux/IAA and 23 ARFs (5, 15 and 3 for A, B and C-ARFs respectively), the NAP of Arabidopsis has evolved into an intricate web of unique interactions, different affinities, and diverse expression domains (Calderón Villalobos *et al.*, 2012; Causier *et al.*, 2012; Dharmasiri *et al.*, 2005; Li *et al.*, 2011; Rademacher *et al.*, 2011; Wang *et al.*, 2013). These duplications likely led to sub- and neofunctionalization of the NAP components. Indeed, the A-ARFs from Arabidopsis (*AtARF5, 6, 7, 8,* and *19*) originated from a single family of A-ARFs that duplicated and diverged into ARF5/7/19 and ARF6/8 subfamilies in ferns **(Fig. 2)**; ARF5, ARF7/19 and ARF6/8 in gymnosperms, and finally five subfamilies in angiosperms (Mutte *et al.*, 2018). Phenotypes of single and double mutants confirm subfunctionalization as *Atarf5* mutants do not form primary roots (Berleth and Jurgens, 1993; Okushima *et al.*, 2005), *Atarf6 Atarf8* mutants have flower defects (Goetz *et al.*, 2006; Nagpal *et al.*, 2005; Tabata *et al.*, 2010; Wu *et al.*, 2006), and *Atarf7 Atarf19* mutants do not produce lateral roots (Okushima *et al.*, 2007; Wilmoth *et al.*, 2005). Notably, the involvement of ARFs in root and flower development, two major evolutionary morphological changes in the tracheophytes and spermatophytes, respectively **(Fig. 2)**, indicate how auxin signalling has been involved in shaping land plant evolution. The most recently duplicated subfamilies (*ARF7/19* and *ARF6/8*) exhibit partial genetic redundancy in Arabidopsis, but these studies nevertheless indicate that the A-ARFs functionally subspecialized, although whether this process additionally involved changes in DNA-binding preferences (neofunctionalization) is unknown. At the centre, understanding how the DNA-binding preferences of the ARFs diverged through evolution is key to understanding how auxin signalling evolved.

M. polymorpha harbours the least complex NAP identified to date and since it did not experience whole genome duplications, its NAP likely represents the ancestral state of the LCA of embryophytes (Bowman *et al.*, 2017; Kato *et al.*, 2020; Mutte *et al.*, 2018). As a haploid species without redundant genes, the NAP of *M. polymorpha* has been extensively investigated to elucidate the basic principles of the NAP. Briefly, MpARF1 transduces the auxin signal by its interactions with MpIAA and thereby switches between repressive and activating activities depending on the auxin concentrations. MpARF2 acts as an attenuator, competing with MpARF1 for similar DNA sites. Cell-specific levels of MpARF1 and MpARF2 creates specific stoichiometries that balance the auxin signal (Kato *et al.*, 2015; Kato *et al.*, 2020). Multiple lines of evidence suggest MpARF3 functions independently of the NAP (Flores-Sandoval *et al.*, 2018a; Flores-Sandoval *et al.*, 2018b; Mutte *et al.*, 2018). Nevertheless, the DNA-binding preferences of the MpARFs have not been investigated in detail. The low redundancy and mirroring of the ancestral state make *M. polymorpha* an ideal model species for investigating the evolution of the ARF DNA-binding preferences.

Scope of the thesis

The DNA-binding preferences of ARFs are critical to understanding how a single molecule can elicit a multitude of responses in plants, and these preferences have therefore been studied extensively *in vitro* from ARFs of angiosperm species. Based on the conservation of the DBD structure and DNA-binding residues, in addition to their capacity to bind AuxRE derived from angiosperm species (e.g., TGTCGG, DR5, IR7) it has been hypothesized that ARFs from other streptophyte lineages (i.e. bryophytes and streptophyte algae) have similar DNA-binding preferences as the angiosperm ARFs. Here, we aim to elucidate the extent to which these ARFs bind similar sequences and how evolvable the ARF-AuxRE interface is.

Chapter 2 reviews and discusses in detail what is known about the DNA-binding preferences of the ARFs and some possible mechanisms that alter or fine-tune their DNA-binding. We discuss recent large-scale DNA-binding assays with angiosperm ARFs that indicate that A- and B-ARFs have different spacing preferences and discuss each motif in detail. Further discussed is how oligomerization and protein degradation could affect DNA-binding and *in vivo* ARF concentrations. Finally, research directions are suggested.

The genomic revolution following the invention of high-throughput sequence has led to massive increases in available sequences. In parallel, the revolutionary protein structure predictor AlphaFold was released. In **Chapter 3** we capitalize on these advances by investigating the conservation of individual residues residues and evolutionary couplings between for the ARF-DBD across the plant kingdom. Furthermore, by analysing the crystal structures of AtARF1 and MpARF2 we updated the ARF DNA-binding interface.

Chapter 4 reports on the DNA-binding preferences of the sole activator ARF in *M. polymorpha*: MpARF1. A bacterial-1-hybrid strategy was used to interrogate binding of MpARF1 on DR5 motifs. By using DNA Affinity Purification followed by high-throughput sequencing (DAP-seq), we report on the AuxRE binding preferences of MpARF1, including spacing preferences for bipartite motifs. These results were supported by a meta-analysis of auxin-treated *M. polymorpha* transcriptomes. Surprisingly, we identified that MpARF1 prefers different bipartite motifs compared to the angiosperm ARFs. Potential binding mechanisms to these motifs were derived by *in silico* structural modelling.

Chapter 5 focuses on the evolvability of the ARF-AuxRE interface by identifying variations for the DNA-binding residues across the plant kingdom. The effects of this

variation were tested *in vivo* by generating MpARF1 variants with altered DNA-binding residues and assessing their effects in a series of phenotyping experiments. Some MpARF1 variants exhibited altered auxin-responses, which were investigated in more detail using transcriptomics. This transcriptomics dataset was additionally used to investigate the role of MpARF1 in the NAP of *M. polymorpha*. Further, *in silico* docking structures were generated to understand the mechanics of altered DNA binding by the MpARF1 variants.

The previous chapters indicated that ARF AuxRE preferences were likely strongly conserved across streptophytes. In **Chapter 6**, we investigated this hypothesis using a series of B3 domain swaps. We discuss the strong conservation of AuxRE preferences for the ARFs in the context of ARF evolution.

Chapter 7 summarizes and discusses the most important finding of this thesis and provides directions for future research.

Chapter

2

To bind or not to bind: how Auxin Response Factors select their target genes

Juriaan Rienstra[1], Jorge Hernández-García[1] and Dolf Weijers[1]

Journal of Experimental Botany, 74, 6922-6932

[1] Laboratory of Biochemistry, Wageningen University and Research, Wageningen, the Netherlands

Abstract

Most plant growth and development processes are regulated one way or another by auxin. The best-studied mechanism by which auxin exerts its regulatory effects is through the nuclear auxin pathway (NAP). In this pathway, AUXIN RESPONSE FACTORs (ARFs) are the transcription factors that ultimately determine which genes become auxin-regulated by binding to specific DNA sequences. ARFs have primarily been studied in *Arabidopsis thaliana*, but recent studies in other species have revealed family-wide DNA-binding specificities for different ARFs and the minimal functional system of the NAP system, consisting in a duo of competing ARFs of the A and B classes. In this review, we provide an overview of key aspects of ARF DNA-binding such as auxin response elements (TGTCNN) and tandem repeat motifs, and consider how structural biology and in vitro studies help us understand ARF DNA preferences. We also highlight some recent aspects related to the regulation of ARF levels inside a cell, which may alter the DNA-binding profile of ARFs in different tissues. We finally emphasize on the need to study minimal NAP systems to understand fundamental aspects of ARF function, the need to characterize algal ARFs to understand how ARFs evolved, how cutting-edge techniques can increase our understanding of ARFs, and which remaining questions can only be answered by structural biology.

Keywords

Auxin, Auxin Response Element, AuxRE, AUXIN RESPONSE FACTOR, dimerization, DNA-binding, gene expression, transcription, transcription factors

Introduction

Auxin is a signaling molecule involved in virtually every plant developmental process (Kou *et al.*, 2022; Weijers and Wagner, 2016). Most responses triggered by auxin are mediated by the nuclear auxin signaling pathway (NAP) **(Fig. 1)**. The NAP is composed of three auxin-specific elements: the TRANSPORT INHIBITOR 1/AUXIN SIGNALING F-BOX (TIR1/AFB) family of receptors, the AUXIN/INDOLE-3-ACETIC ACID (Aux/IAA) family of corepressors, and the AUXIN RESPONSE FACTOR (ARF) family of DNA-binding transcription factors. Auxin acts as a molecular glue to promote interactions between TIR1/AFB and Aux/IAA proteins, leading to the ubiquitination and subsequent destabilization of the latter (Tan *et al.*, 2007). Aux/IAA proteins act as transcriptional corepressors by directly interacting with ARFs and imposing transcriptional repression (Tiwari, 2001; Ulmasov *et al.*, 1997b). Upon Aux/IAA degradation, ARFs are released to regulate gene expression (Tiwari, 2001; Tiwari *et al.*, 2003). In the end, the ARFs are the ultimate arbiters that decide which genes become auxin-regulated by virtue of their sequence-specific DNA-binding (Ulmasov *et al.*, 1997b).

ARF proteins contain three conserved domains with distinct functions: an N-terminal DNA-binding domain (DBD), a middle region (MR), and a C-terminal Phox and Bem1 (PB1) domain **(Fig. 1)** (Tiwari *et al.*, 2003; Ulmasov *et al.*, 1999b; Weijers and Wagner, 2016). The DBD contains a B3 domain that allows ARFs to bind DNA motifs called Auxin Response Elements (AuxREs) (Boer *et al.*, 2014; Ulmasov *et al.*, 1999b; Weijers and Wagner, 2016). Other regions within the DBD facilitate homodimerization and heterodimerization, imposing an additional layer of DNA-binding preference towards tandem repeat AuxREs (see further) (Boer *et al.*, 2014; Ulmasov *et al.*, 1997a, 1999b). Once bound to chromatin, ARFs can either repress or activate transcription, through their MR (Kato *et al.*, 2015; Ulmasov *et al.*, 1999a). The PB1 domain acts as a homo- and heterotypic oligomerization domain, allowing ARFs to homo- and/or hetero-oligomerize and/or interact with Aux/IAA proteins (Hamann *et al.*, 2002; Kato *et al.*, 2015; Kim, 1997; Li *et al.*, 2011; Ouellet *et al.*, 2001).

ARFs likely originated in a streptophyte common ancestor to land plants and algal sisters, and diverged into three phylogenetically and functionally separated classes present in all subsequent land plants: A, B, and C (Mutte *et al.*, 2018). Extant streptophyte algal ARFs belong either to the C-class, or to an A/B-class that likely represents an ancestral state preceding A/B diversification (reviewed in (Carrillo-Carrasco *et al.*, 2023; Kato *et al.*, 2018). While algal ARFs have not been studied in depth, analysis of the single ARF in *Chlorokybus atmophyticus* suggest a DNA-binding preference similar to that of land plant counterparts (Martin-Arevalillo *et al.*, 2019). This indicates that the fundamental ARF DNA-binding preferences are likely ancestral and have been maintained during evolution. In land plants, ARF transcriptional activities are well conserved between bryophytes and angiosperms and are attributable to the

different ARF classes: A-ARFs are known to act mainly as transcriptional activators, while B- and C-classes function as repressors (Kato *et al.*, 2015; Tiwari *et al.*, 2003). A- and B-class ARFs are known to regulate auxin transcriptional responses in land plants, whereas the integration of C-ARFs into the auxin response network is less clear (Flores-Sandoval *et al.*, 2018a; Flores-Sandoval *et al.*, 2018b; Mutte *et al.*, 2018). Among land plants, the genome of the liverwort *Marchantia polymorpha* encodes the simplest NAP studied to date, with a single TIR1/AFB, a single Aux/IAA, and a single ARF for each of the three classes (Bowman *et al.*, 2017; Flores-Sandoval *et al.*, 2015; Kato *et al.*, 2015; Kato *et al.*, 2017; Kato *et al.*, 2020; Mutte *et al.*, 2018). In contrast, repeated gene and genome duplications have shaped more complex NAP families in flowering plants, represented

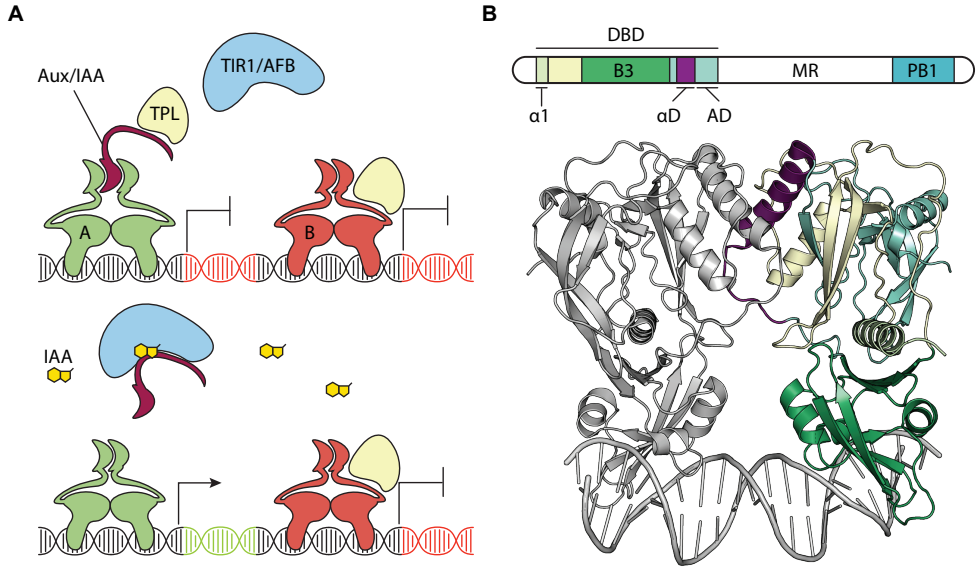


Fig. 1 | Overview of the Nuclear Auxin Pathway and the anatomy of an ARF. (A) The minimalistic nuclear auxin pathway. Under low auxin conditions (upper panel), both A-ARFs (green) and B-ARFs (red) act as repressors. A-ARFs bind with a repressive cofactor, Aux/IAA (bordeaux), that recruits TPL (yellow), while B-ARFs recruit TPL directly via their middle region. Under increasing auxin conditions (indole-3-acetic acid, IAA, gold), IAA acts as a molecular glue and allows TIR1/AFB (blue) to sequester Aux/IAA away from the A-ARFs, which then become transcriptional activators. B-ARFs act as repressors in either condition. **(B)** The anatomy of an ARF. Top panel shows the general genetical sequence of an ARF, the lower panel shows the atomic structure of the DNA-binding domain in complex with an IR7 motif (PDB: 6ycq). Domains indicated and discussed in this review are: α1 (pear), the alpha-helix tethering the B3 domain and acting as a molecular hinge; N-terminal dimerization domain (yellow); B3 (green), the domain interacting with the DNA; αD (purple), the alpha-helix and loop that facilitates dimerization; and the C-terminal dimerization domain and ancillary domain (AD, cyan). The middle region (MR, white) and Phox and Bem1 domain (PB1, blue) are omitted.

by six TIR1/AFB, 29 AUX/IAA and 23 ARF (5A, 15B, 3C) members in *Arabidopsis thaliana* (Arabidopsis) (Mutte *et al.*, 2018; Weijers and Wagner, 2016). In *M. polymorpha*, the single A-ARF and B-ARF (MpARF1 and MpARF2 respectively) compete for the same DNA motifs (Kato *et al.*, 2020), a mechanism shared with representatives of the moss *Physcomitrium patens* A- and B-class ARFs (Lavy *et al.*, 2016). In this model, MpARF1 is an auxin-sensitive transcriptional regulator that switches between Aux/IAA-imposed transcriptional repression in the absence of auxin, and ARF-mediated transcriptional activation in the presence of auxin, while MpARF2 represses independently of auxin and acts to limit auxin responsiveness **(Fig. 1)**. This basic module may reflect the system present in the (elusive) last common ancestor of land plants. However, this module expanded significantly during land plant evolution, leading to an increase in the web of interactions, to redundancy between ARFs, and sub- and neo-functionalization of ARFs (Hardtke *et al.*, 2004; Li *et al.*, 2006; Nagpal *et al.*, 2005; Okushima *et al.*, 2005; Rademacher *et al.*, 2012; Rademacher *et al.*, 2011; Weijers *et al.*, 2005).

In this review, we will discuss large-scale DNA-binding assays and promoter studies that have elucidated ARF binding specificities. We will discuss this in the context of ARF DNA-binding structural biology, and then briefly focus on other aspects that likely affect ARF DNA-binding such as oligomerization, cytoplasmic condensate formation, and ARF E3 ligase-mediated degradation. We touch on aspects that add levels of complexity to the basic A/B competition model proposed for *M. polymorpha*, and that may underlie the intricate scenario of auxin transcriptional regulation in other land plants.

Selective ARF binding to auxin-response elements

ARFs select their target genes based on the presence of AuxREs, defined as TGTCNN elements (where N can be any nucleotide). One of the first identified AuxREs was the TGTCTC motif of the soybean *GRETCHEN HAGEN 3* (*GH3*) promoter, later used as a "bait" to identify the first ARF: AtARF1 (Liu *et al.*, 1994; Ulmasov *et al.*, 1997a, 1999b; Ulmasov *et al.*, 1995). The TGTC acts as the invariable core element vital for auxin response, while the final two nucleotides are variable (Boer *et al.*, 2014; Liu *et al.*, 1994; Ulmasov *et al.*, 1999b; Ulmasov *et al.*, 1995). Protein binding microarray analyses of Arabidopsis ARFs show that ARFs bind a multitude of TGTCNN elements, each with different affinities, with the highest affinity towards the TGTCGG element (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014). In line with this, a gene expression reporter based on a multimer of TGTCGG has about a tenfold higher auxin sensitivity than a TGTCTC-based reporter (Liao *et al.*, 2015). Recently, genome-wide ARF DNA-binding studies of several maize A- and B-class ARFs indicated that all family members can bind a variety of TGTCNN elements but strongly prefer TGTCGG elements (Galli *et al.*, 2018).

The mechanism for both the generic ability to bind the TGTC core and the

preference for TGTCGG elements can be found in the crystal structure. The DNA-binding residues were first identified in AtARF1 in complex with TGTCTC DNA (4ldx; Boer *et al.*, 2014)**(Fig. 1)**. More recently, the structures of AtARF1 (6ycq; Freire-Rios *et al.*, 2020) and MpARF2 (6sdg; Kato *et al.*, 2020) in complex with TGTCGG DNA was solved. This allowed us to re-examine and update the residues involved in DNA-binding using the AtARF1 sequence as reference **(Fig. 2)**. The B3 domain harbours two types of DNA-binding residues, the first off which are residues that interact with the DNA phosphate backbone to confer general DNA-binding affinity via hydrogen bonds and electrostatic interactions **(Fig. 2A, 2C)**. The second type of DNA-binding residues enter the DNA major groove and confer TGTCNN specificity. These are found in two loops: a C-terminal loop of which residues Q183, P184, R181 and R186 bind the core TGTC and an N-terminal loop where G137 and H136 are involved in binding the two variable residues **(Fig. 2A ,**

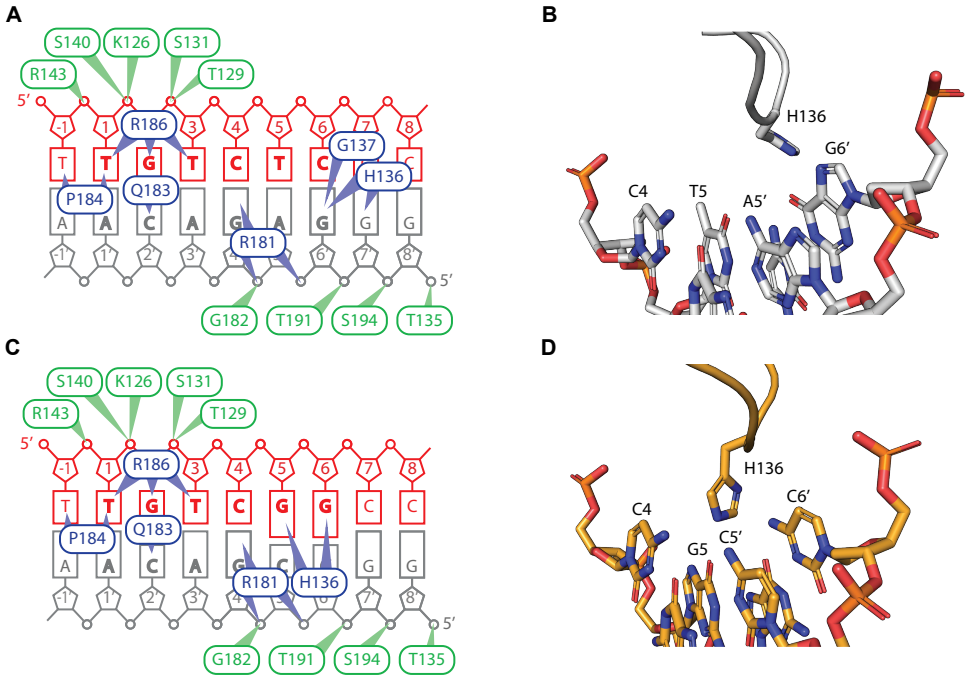


Fig. 2 | Atomic interface between AuxRE and ARF-B3. Residues are numbered according to AtARF1, DNA bases are numbered according to the first nucleotide of TGTCTC. **(A, C)** Schematic representation of DNA contact sites of ARF-B3 for TGTCTC (A) and high affinity TGTCGG (C). Contacting residues are indicated with ovals, green for residues only contacting the phosphate backbone and blue for residues contacting specific DNA bases. The figures are composites of three structures (4ldx, 6ycq, 6sdg; cutoff for contacts of 3.5 Å) and as such, some contacts are only found in a single structure, or compensated for by other residues. For instance, either K126 makes contacts alone, or T129 and S140 contact together, but never all three at the same time. **(B, D)** Closeup of AtARF1-H136 conformation change with AuxRE bases of either TGTCTC (B) or TGTCGG (D). Based on Freire-Rios *et al.* (2020).

2C). TGTCGG elements are likely bound with higher affinity than TGTCTC because H136 is rotated deeper into the major groove and forms additional hydrogen bonds with G5 and G6 the TGTCGG structure **(Fig. 2C, 2D)**. In comparison, in the TGTCTC structure this histidine is interacting only with G6' and G7', and is sterically hindered from rotating deeper into the DNA by C6 and G6' **(Fig. 2A, 2B)** (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020). It is therefore likely that this H136-containing loop is responsible for the wide variety of TGTCNN elements that ARFs can bind depending on the rotamer and steric hindrances. Given that the general DBD structure and the DNA major groove binding residues are strongly conserved across ARFs from the plant kingdom (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Kato *et al.*, 2020; Mutte *et al.*, 2018), it is likely that TGTCGG elements were the preferential DNA-binding elements for the A- and B-class ancestors. This would also support the notion of A/B competition being an ancestral characteristic of ARF-mediated transcriptional regulation in land plants.

While ARFs show a clear preference for TGTCGG elements, these are not pervasive in auxin-responsive genes and instead, auxin-responsive promoters have a wide variety of TGTCNN elements, (Cherenkov *et al.*, 2018; Donner *et al.*, 2009; Lieberman-Lazarovich *et al.*, 2019; Liu *et al.*, 1994; Walcher and Nemhauser, 2012; Xu *et al.*, 1997; Zemlyanskaya *et al.*, 2016). Still, some of the different TGTCNN elements can be associated with particular modes of auxin response. For example, early auxin-activated genes are associated with TGTCTC elements, such as those found in the *GH3* and *SMALL AUXIN UPREGULATED RNAs* (*SAUR*) promoters, while the high-affinity TGTCGG element is found in both early and late auxin-activated and -repressed responses (Zemlyanskaya *et al.*, 2016) (Cherenkov *et al.*, 2018; Lieberman-Lazarovich *et al.*, 2019). Elements can also be associated with specific developmental or cellular processes, such as TGTCGG with cell-wall-related genes or TGTCTT with phosphate metabolism, the latter which is shown to likely involve AtARF7 and AtARF19 (Huang *et al.*, 2018; Lieberman-Lazarovich *et al.*, 2019; Narise *et al.*, 2010; Zemlyanskaya *et al.*, 2016). Finally, these elements can be associated with different expression domains. For instance, in the Arabidopsis roots, the same gene expression reporter with different TGTCNN elements has been tested, of which TGTCGG and TGTCTC differ in the location of expression maxima, with TGTCTC reaching a maximum in the quiescent centre, and TGTCGG in the subtending columella cells (Liao *et al.*, 2015). Reporters with TGTCCC, TGTCGC, and TGTCAC show both overlapping and unique expression domains (Lieberman-Lazarovich *et al.*, 2019). These unique domains suggest that they are caused by particular ARFs with alternative DNA-binding preferences, but so far, it is unknown whether any correlation exist between different TGTCNN reporters and the unique ARF expression domains in the roots (Lieberman-Lazarovich *et al.*, 2019; Rademacher *et al.*, 2011).

Higher-order DNA-binding specificity through ARF dimerization

Few studies have identified ARF targets (see below), but those available suggest distinct sets of targets. However, if all ARFs preferentially bind the same element, TGTCGG, it would be difficult to explain how ARFs regulate distinct sets of target genes. Rather, like many eukaryotic transcription factors (Amoutzias *et al.*, 2008; Kribelbauer *et al.*, 2019; Sloan *et al.*, 2020), ARFs bind DNA as dimers, and can homodimerize via their DBD (Boer *et al.*, 2014), and additionally through their PB1 domain (Han *et al.*, 2014; Kim *et al.*, 2020; Korasick *et al.*, 2014; Nanao *et al.*, 2014), to bind tandem repeat motifs of TGTCNN elements. The nomenclature of these tandem repeat is as set by Freire-Rios *et al.* (2020), with Inverted Repeat (IR), Everted Repeat (ER) and Direct Repeat (DR) depending on the orientation of two TGTCNN elements and the number of bases between the individual motifs **(Fig. 3A)**. For example, when two TGTCNN elements point to the same direction (direct repeat) and are spaced five nucleotides apart, this is a DR5 motif. Given that A-ARFs are auxin-dependent activators and B-ARFs are auxin-independent repressors (Kato *et al.*, 2015; Kato *et al.*, 2020; Tiwari *et al.*, 2003), having unique motif preferences could dictate certain target genes free from A/B competition, and are therefore either generally activated or repressed.

Early work identified DR5 and IR7 motifs to be effective auxin-inducible reporters (Ulmasov *et al.*, 1997a; Ulmasov *et al.*, 1997b). Recently, large-scale DAP-seq studies have yielded the first comprehensive overview of genome-wide DNA-binding by ARFs, revealing both overlapping and distinct motif preferences for A-ARFs and B-ARFs (Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). In both Arabidopsis and maize, A- and B- ARFs can bind IR7/8 motifs, while A-ARFs are additionally capable of binding to several DR and ER motifs. While there are stark differences between the DNA-binding specificities of A- and B-classes of ARFs, relatively minor differences exist within each class. One caveat is that in DAP-seq, transcription factors are offered genomic DNA fragments devoid of histones, leading to binding events also occurring within regions normally inaccessible in a physiological context, highlighted by a low overlap of binding events between DAP-seq and ChIP-seq experiments (O'Malley *et al.*, 2016). In maize, only 5-25% of DAP-seq ARF binding events occur within open chromatin regions (Galli *et al.*, 2018). However, when combining DAPseq with gene expression data in maize, it appeared that auxin-activated genes either have a sole A-ARF binding event, or shared A-ARF and B-ARF events, but never a B-ARF-only event (Galli *et al.*, 2018). This matches with data obtained in Arabidopsis, where promoters of auxin-activated genes are enriched for both binding events and motifs that are targeted primarily by A-ARFs (see below) (Freire-Rios *et al.*, 2020; Stigliani *et al.*, 2019). This supports the idea that genes are either auxin-activated or -repressed as a function of which ARFs can bind their regulatory regions. Hence, the different conformations of tandem elements represent

an additional layer of complexity to ARF-mediated transcriptional regulation.

Inverted Repeats

The IR7/8 conformation likely represents the canonical ARF competition model observed in *M. polymorpha*: A- and B-ARFs from *M. polymorpha* and Arabidopsis are able to bind this conformation, although B-ARFs bind with slightly less affinity than A-ARFs **(Fig. 3B)**, consistently, both auxin-activated and auxin-repressed genes in Arabidopsis are associated with these motifs (Boer *et al.*, 2014; Fontana *et al.*, 2023; Freire-Rios *et al.*, 2020; Kato *et al.*, 2020; Stigliani *et al.*, 2019). In agreement, DAP-seq binding events containing these motifs are enriched when the events are filtered by auxin-responsive genes, both in maize and Arabidopsis (Galli *et al.*, 2018; Stigliani *et al.*, 2019). Single-molecule in vitro ARF-DNA binding assays suggest two different modes of binding: for B-class ARFs (AtARF1 and AtARF2), both the DBD and PB1 domains facilitate cooperative homodimerization. Monomeric ARFs have a low affinity for DNA, and quickly dimerize via the DBD and PB1 domains. Once dimerized, they bind IR7 tandem motifs forming a more stable complex. However, this stabilization relies on the cooperative action of the DBD and PB1, as removing the PB1 leads to AtARF2 being primarily present as monomers in solution (Fontana *et al.*, 2022; Fontana *et al.*, 2023). This dual mode of cooperative binding likely holds true for other ARFs as well (Ulmasov *et al.*, 1999b). For the A-class AtARF5, even though the AtARF5-DBDs are present primarily as monomers in solution, the DBD itself is sufficient for strong dimerization and DNA-binding, relying less on the cooperative action of the PB1 domain (Fontana *et al.*, 2022; Fontana *et al.*, 2023). ARF5 PB1 domain has a higher affinity towards Aux/IAA than to homodimerization (Fontana *et al.*, 2023; Hamann *et al.*, 2002; Kim, 1997; Kim *et al.*, 2020; Li *et al.*, 2011; Muto *et al.*, 2006; Ouellet *et al.*, 2001), which matches with its primary function of being an auxin-dependent (through recruiting Aux/IAA) transcriptional activator (Tiwari *et al.*, 2003). Physiological evidence supports the importance of cooperative dimerization for proper ARF functioning: *AtTMO5* (itself a direct target of AtARF5) and *AtIAA19* promoters each have an IR motif that is only functional if an ARF dimer binds (Freire-Rios *et al.*, 2020). In line with this, *AtARF5* mutants in the dimerization domain of the DBD are unable of rescuing an *Atarf5/mp* mutant (Boer *et al.*, 2014).

The crystal structures of AtARF1 (4ldx, 6ycq; (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020)) and MpARF2 (6sdg; (Kato *et al.*, 2020)) DBDs revealed the binding mode on IR7 motifs. The DBD structure can be divided in three subdomains **(Fig. 1B)**. The B3 domain, discussed above, is tethered by a single alpha helix (α1) underneath a fold composed of the other two subdomains: the dimerization domain (DD) and the ancillary domain (AD). The DD face is composed of an alpha helix (αD) that facilitates dimerization with the mirrored helix αD of the other monomer, while simultaneously a dimerization loop is inserted into a groove of the opposite monomer, leading to dimerization in a head-to-head orientation. The AD is genetically located at the C-terminal end of the DBD, and

occupies the tail position in the structure of the dimer (Boer *et al.*, 2014). Upon dimer binding to the IR7 DNA, the DNA is slightly contorted and the B3 domains twist on helix α1 to enable placement of the two DNA-binding loops (discussed above). Even though AtARF1 and MpARF2 have had 500 million years of independent parallel evolution, their DNA-bound, dimerized structures are essentially the same (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Kato *et al.*), suggesting a strong evolutionary pressure to maintain the ARF fold as it is.

Besides IR7/8, other IR motifs are likely selected by ARFs but are possibly rare. In carrot cells, promoters with IR7/8 are strongly auxin-inducible, while IR5/6/9 are weakly inducible (Ulmasov *et al.*, 1997a). In SPR analysis, AtARF1 (B) was found to be restricted to IR7/8 motifs, while AtARF5 (A-class) could bind IR5, IR6 or IR9 motifs, although with less affinity than IR7/8 (Boer *et al.*, 2014). Given how the B3 domains twist on α1 to adapt to IR7 DNA, it is plausible that, compared to B-ARFs, more intrinsic flexibility within A-ARFs in the helix α1 region could accommodate to different IR conformations other than IR7/8. This model was termed the 'caliper model' by Boer *et al.* (2014)**(Fig. 3C)** and could offer an explanation why ARFs cannot functionally complement each other (Hardtke *et al.*, 2004; Kato *et al.*, 2020; Li *et al.*, 2006; Nagpal *et al.*, 2005; Okushima *et al.*,

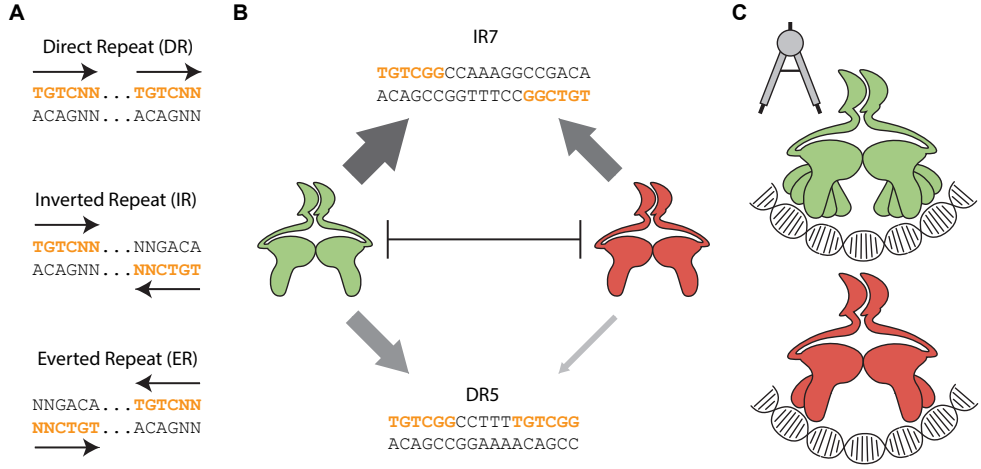


Fig. 3 | Spacing motifs determine ARF specificity. (A) Definition of direct repeat (DR), inverted repeat (IR) and everted repeat (ER), following definition set by Freire-Rios *et al.* (2020). N denotes A, T, C or G, while dots represent variable numbers of intermittent nucleotides between two TGTCNN elements. **(B)** Schematic figure representing single-molecule FRET analysis of DNA-binding. Arrow thickness represents affinity. A- (green) and B-ARFs (red) are competing for DNA-binding. Both ARFs have an affinity for IR7 motifs that allows for in vivo competition on the same locus. For DR5, both ARFs have a weaker affinity, with B-ARF affinity being so weak, their in vivo binding is likely negligible (Freire-Rios *et al.*, 2020; Kato *et al.*, 2020). **(C)** The proposed caliper model (Boer *et al.*, 2014). The A-ARF AtARF5 (green) was found to be able to bind IR5 to IR9 motifs, whereas the B-ARF AtARF1 (red) was limited to IR7/8.

2005; Rademacher *et al.*, 2012; Rademacher *et al.*, 2011; Weijers *et al.*, 2005). However, whether the caliper model extends beyond AtARF1 and AtARF5 is unknown. So far, in vivo only IR9 motifs have been identified so far, in promoters of *AtLFY*, *AtTMO3* and *AtIAA19* (Freire-Rios *et al.*, 2020; Rademacher *et al.*, 2011; Yamaguchi *et al.*, 2013). It is therefore likely that when it comes to IR motifs, motifs other than IR7/8 are likely functional, but since they are not identified with either DAP-seq or promoter meta-analyses (Freire-Rios *et al.*, 2020; Galli *et al.*, 2018), these events are likely rare.

Direct Repeats

DR motifs appear to be conformations that are relatively specific to A-ARFs and auxin-dependent activation (Freire-Rios *et al.*, 2020; Galli *et al.*, 2018; Stigliani *et al.*, 2019). DR5 became the *de facto* auxin response reporter based on its strong auxin-inducibility (Ulmasov *et al.*, 1997b), well before ARF motif conformation spacing was studied and DR5 motifs were identified in auxin-regulated genes (Lieberman-Lazarovich *et al.*, 2019). A possible explanation for why DR5 is an effective reporter can be found in the most recent DAP-seq data, where only A-ARFs appear to bind DR4/5 motifs (Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). In maize, auxin-activated genes are associated with A-ARF binding events, and in Arabidopsis, auxin-activated genes are associated with AtARF5 binding to DR5 motifs (Freire-Rios *et al.*, 2020; Galli *et al.*, 2018). These results suggest that only A-ARFs and not B-ARFs are capable of binding to DR5 in vivo. In vitro, however, both classes are able to bind DR5, albeit with much weaker affinities than to IR7 or IR8 motifs **(Fig. 3B)**(Freire-Rios *et al.*, 2020; Kato *et al.*, 2020). If we speculate, it might be possible that B-ARF affinities are so low that DR5 occupancy by B-ARFs is negligible or that B-ARFs are titrated from DR5 sites by high-affinity (IR7/8) sites, such that A-ARFs can more easily bind DR5 motifs and effectively turn them into a net strong auxin-activated motif (Freire-Rios *et al.*, 2020). Structurally, we can only speculate how DNA-binding is achieved. A DR5 motif composed of two TGTC repeats becomes auxin-unresponsive when either element is mutated (Liu *et al.*, 1994; Ulmasov *et al.*, 1997b), suggesting that a dimer binds cooperatively. Furthermore, binding analyses were performed with ARF DBDs only, indicating that the DBD is sufficient for binding DR5 motifs (Freire-Rios *et al.*, 2020). The 5-bp spacing in DR5 is a full DNA-helix twist suggesting that the dimer binds on the same plane. This would place the DBDs in a head-to-tail orientation, that is, with the dimerization α-helix (αD) and AD **(Fig. 1B)** potentially interfacing each other. DR4 and DR6 would place DBDs slightly off kilter, are only weakly auxin-inducible in carrots (Ulmasov *et al.*, 1997b), and only DR4 is bound by maize and Arabidopsis A-ARFs in vivo (Galli *et al.*, 2018; Stigliani *et al.*, 2019). All in all, while the DR5 reporter celebrates its 25th anniversary, the motif conformation still harbors secrets.

Everted Repeats

ER motifs remain an enigma, as no functional sites have been identified in vivo nor its

structure has been solved, yet ER motifs are enriched in DAP-seq experiments. ER motifs are only weakly enriched in promoters of auxin-regulated genes (Freire-Rios *et al.*, 2020). Still, both A- and B-ARFs can bind ER motifs in vitro, although their spacing preference is family-specific (Galli *et al.*, 2018; Stigliani *et al.*, 2019). For AtARF5, an ER0 motif was found to be enriched, which would have one of the monomers on the complete 180° opposite side of the DNA helix. This orientation would preclude dimerization via the DBD for the ARFs, therefore relying on the PB1 domain, to facilitate DNA-binding. AtARF5 also binds ER13, placing the ARFs on the same plane of the DNA helix, in a tail-to-tail (AD to AD) conformation (Stigliani *et al.*, 2019). Given that ER motifs are enriched in DAP-seq, it at least indicates the presence of such conformations in the genome but their functional importance is yet to be proven in vivo.

Control of ARF localization and stability

ARFs are the pivot point in auxin-dependent gene regulation. As discussed above, intrinsic properties provide DNA binding specificity, while Aux/IAA association determines activity and A/B competition defines auxin sensitivity. Within this relatively simple response system, behavior is dictated by biochemical interaction parameters. Biological interactions though are a function of both affinity and concentration, and therefore, any mechanisms that control (local) protein concentration can impact ARF action (Kribelbauer *et al.*, 2019).

ARF PB1 crystal structures revealed that these domains can oligomerize (Han *et al.*, 2014; Kim *et al.*, 2020; Korasick *et al.*, 2014; Nanao *et al.*, 2014). The PB1 domain is composed of a positively charged face (mainly due to an invariant lysine), and a negatively charged face (named OPCA after OPR, PC, and AID) (Ponting *et al.*, 2002). Two PB1 domains will interact in a head-to-tail manner, and more PB1 domains could oligomerize on either side. Oligomerization is clearly a mechanism to create higher local protein concentration, perhaps to enhance binding to low-affinity sites, or to cause transcriptional bursts (Kribelbauer *et al.*, 2019; Wei *et al.*, 2020). However, it is not yet clear if ARF oligomerization enhances DNA-binding, as additional oligomers do not enhance the binding on a single IR7 motif (i.e. two TGTCGG elements) in vitro (Fontana *et al.*, 2023), although oligomerization does enhance activation in a heterologous yeast system (Pierre-Jerome *et al.*, 2016). In DAP-seq experiments, both A-ARFs and B-ARFs bind to motifs that have a spacing that is too large for a single dimer to bind, and it was proposed they are bound by oligomers (Galli *et al.*, 2018; Stigliani *et al.*, 2019). Outside of the nucleus, oligomerization appears to function in sequestering A-ARFs into the cytoplasm: AtARF19 accumulates in oligomerization-driven cytoplasmic condensates, in particular in differentiated root cells. The condensation appears to involve the intrinsically disordered – prion-like – Middle Region, but critically requires interactions through the PB1 domain. Mutating the charged faces of the AtARF19 PB1

domain restores nuclear localization, prevents condensation, and leads to hyperactivity of the ARF (Powers *et al.*, 2019). Whether oligomerization is involved in DNA-binding, controlling nuclear levels of ARFs, or if it serves a different function will be questions for the future.

Genome-wide, TGTCNN elements are widespread, yet only a fraction of them are bound by ARFs. ARF dimerization is one way to add specificity for DNA binding, but like any equilibrium, monomeric species exist in the cell nucleus with an inherent capability to bind single TGTCNN motif, that would be expected to occur once every 256 basepairs by chance. Recently, it has been suggested that some of these monomeric forms are actively degraded to prevent monomeric binding that could trigger undesired transcriptional changes. In *M. polymorpha*, both MpARF1 and MpARF2 are degraded via the proteasomal machinery. The MpARF2 degron motif overlaps with the dimerization helix (αD, **Fig. 1B**) of the DBD, and is likely unavailable upon dimerization, thus preventing degradation of dimeric MpARF2. Non-dimerizing mutants are rapidly degraded, but when these monomeric species are additionally mutated in the degron motif, the resulting plants show serious defects in thallus development (Das *et al.*, 2022). This suggests that monomeric species can cause significant transcriptional misregulation in the cell, and degradation of these monomeric species could be a manner to shift the equilibrium to dimeric species.

There is a precedent for ARF protein degradation. AtARF1 was shown to be an unstable protein with a half-life of about 3 hours (Lakehal *et al.*, 2019; Salmon *et al.*, 2008). This instability was shown to depend on proteasomal degradation, but not on the Skp, Cullin, F-box containing complex (SCF) subunit CULLIN 1. The biological relevance of degradation was not yet established, but there clearly is scope for active control of AtARF1 accumulation. Likewise, AtARF2 appears to be subject to proteasomal degradation: the hormone GA promotes AtARF2 accumulation, likely by preventing its degradation (Richter *et al.*, 2013). AtARF1, AtARF6, AtARF8 and AtARF17 are all actively degraded via the proteasome in Arabidopsis protoplasts (Lakehal *et al.*, 2019). More recently, it was reported that the E3 ligase AFF1 mediates degradation of AtARF7 and AtARF19 in the Arabidopsis root, acting as a mechanism for fine-tuning ARF levels within a cell (Jing *et al.*, 2022). Mutants in the E3 ligase have a weak but significant phenotype in auxin response, which suggests a biological role for ARF degradation. Thus, with these examples from Arabidopsis A-class (AtARF7/19) and B-class (AtARF1/2) and both A- and B-class *M. polymorpha* ARFs, it is emerging that ARF degradation may be a common principle. This raises several intriguing questions. Are all ARFs degraded via E3-ligase mediated degradation? Are there multiple mechanisms? Where are the degron motifs located? The biological importance of ARF degradation has been suggested in both *M. polymorpha* and Arabidopsis and will likely lead to many exciting findings.

Open and pressing questions

Recent work in *M. polymorpha* has revealed what is likely the minimal, ancestral-like state of auxin-regulation by the ARFs, wherein the A-ARF is auxin-sensitive and the A- and B-ARFs are competing for the same motifs. It is then the stoichiometry of the two types of ARFs that dictates the auxin-sensitivity of a cell. However, no matter the simplicity of the *M. polymorpha* system, there are still many questions that remain to be answered. Given that, compared to B-ARFs (only IR motifs), angiosperm A-ARFs can likely bind a wider variety of tandem repeat motifs (DR, ER and IR), does this also occur in MpARF1 and MpARF2? Does MpARF1 have a caliper capacity like AtARF5 to extended IR motifs? How much overlap exist between MpARF1 and MpARF2 DNA-binding sites *in vivo*? Answering these questions will provide the basic framework for how A- and B-ARFs work at the simplest level and will allow us to investigate the unique specificities that ARFs acquired during sub- and neo-functionalization of the paralogs existing in angiosperms.

ARFs most likely originate in a common ancestor to all streptophytes, but we do not know what function or characteristics they had. Only one algal ARF has been studied for its biochemical properties, and while it can bind TGTCNN motifs, it suggests its spacing preferences are related to C-ARFs rather than to the A/B-ARFS that it is phylogenetically more related to (Carrillo-Carrasco *et al.*, 2023; Martin-Arevalillo *et al.*, 2019). It will be interesting to find out if the DBD domains of algal ARFs share some or all their properties with land plant ARFs. Do these ARFs dimerize? Did the ancestral ARF have features of A-, B-, and/or C-ARFs or are their characteristics entirely unrelated to land plant ARFs? ARFs were already separated into three clades in the last land plant common ancestor, with *M. polymorpha* having a minimal NAP system that could be reminiscent of how an ancestral system would have functioned. Having such a simple, yet essential system would place evolutionary constraints on mutations that would disrupt the carefully balanced stoichiometry of MpARF1 and MpARF2. That begs the question, can we find other bryophytes with a similar minimal system, and do they share the same features as the ARFs in *M. polymorpha*?

ARF families expanded following whole genome duplications in seed plants, freeing them from the evolutionary constraints of the minimalistic system and allowing for neo- and subfunctionalization. In Arabidopsis and maize, ARFs have unique expression domains, not all ARFs are able to replace each other, and the more phylogenetic distance between ARFs, the less overlap in genomic binding profiles they have. The opposite can also be true: more closely related ARFs act redundantly (AtARF7 and AtARF19) and share more similar DNA-binding patterns (ZmARF4 and ZmARF29). These additional complexities make studying ARFs in flowering plants challenging. Despite relatively good understanding of the generic mechanisms underlying auxin response, there is a pressing scarcity of insight into the actions of individual ARFs in their functional context. Improvements in single-cell resolution methods for DNA-binding

(CUT&RUN) (Skene and Henikoff, 2017), (single-cell) transcriptomics (Kolodziejczyk *et al.*, 2015; Lavrekha *et al.*, 2022), and protein-protein interactions (proximity-labelling techniques) (Cortal *et al.*, 2021) may offer the resolution needed to study ARF function at the granularity necessary to dissect their unique functions in growth and development. In parallel, further improvements in statistical modelling (Lai *et al.*, 2019), and machine learning tools (Alipanahi *et al.*, 2015), should enable us to not only identify the major patterns in the data (IR7/IR8, TGTCGG), but also the rare events (IR9, TGTCTC) that are likely highly relevant in specific contexts, as seen from unique expression domains of different TGTCNN reporters in Arabidopsis roots. These methods would enable us to answer questions such as: what are the ARF caliper potentials? Even though all ARFs favor TGTCGG elements, do their preferences differ for medium affinity elements? What motifs are enriched in auxin-responsive genes on a single-cell level?

ARF DBD and PB1 crystal structures have enabled linking functions to structures for dimerization, DNA-binding, and degradation. However, many questions remain: how do A-ARFs bind DR5? Can crystal structures of A-ARFs with IR5 or IR9 explain the caliper model? And regarding the TGTCNN specificity, while the DNA-binding residues are conserved, there is interspecific variation in the B3 domain itself between ARF paralogs and orthologs. Can this variation affect DNA-binding? Or would all B3 domains target essentially the same motifs?

Over the past three decades, the scientific community has unraveled a simple nuclear auxin signaling pathway that derives into ARFs deciding which genes need to be auxin-regulated. This simplicity is deceptive, as each ARF seems to have has its own peculiarities. Nevertheless, many secrets of the ARFs remain to be revealed, a challenge our research community is ready to tackle.

Acknowledgements

The authors would like to thank Gül Hatinoğlu, Martijn de Roij and Sjoerd Woudenberg for helpful discussion and careful reading of the manuscript. This work was supported by funding from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (grant no. GSGT.GSGT.2018.013) to J.R. and by a Marie Skłodowska-Curie Individual Fellowship (H2020-MSCA-IF-2020) to J.H.-G.

Chapter

3

Conservation and diversity in sequence and structure of the ARF transcription factors

Juriaan Rienstra[1], Jorge Hernández-García[1] and Dolf Weijers[1]

[1] Laboratory of Biochemistry, Wageningen University and Research, Wageningen, the Netherlands

Abstract

The AUXIN RESPONSE FACTOR (ARF) family originated in the Streptophyte lineage, with the last common ancestor harboring an A/B-like and C-like ARF, two classes that persist in extant Charophytes. In terrestrial plants, the A/B-ARFs duplicated and diverged, giving rise to three classes: A-ARFs, B-ARFs and C-ARFs. Despite millions of years of independent evolution between lineages and classes, the ARF protein structure and DNA-binding preferences have been conserved, although a link with conserved residues has remained unknown. The ARF DNA-binding domain (DBD) consists of a B3 domain responsible for DNA interaction, and a dimerization domain (DD) facilitating homodimerization of two ARF monomers, enabling binding to bipartite DNA elements. These domains are interconnected by a flexible molecular hinge (DD-B3 hinge). In this chapter, our goal was to unravel the relationship between structural fidelity and conservation of individual residues. From a reanalysis of the high resolution AtARF1 structure, we identified water-mediated hydrogen bonds and π-interactions extend the known DNA-binding interface. By combining analyses of existing protein structures, sequence conservation data, and evolutionary couplings, we found that both the structural and DNA-binding residues of the B3 domain are extremely conserved, and that several DNA-binding residues are evolutionary coupled with other DNA-binding residues. Furthermore, we revealed that the buried, structural residues of the DD are more conserved compared to surface residues. AlphaFold2 predictions of the DBDs show a high degree of structural similarities across all ARF classes and Streptophyte lineages. Lastly, we found substantial variability within the (homo)dimerization interface and part of the DD-B3 hinge, implying potential differences in dimerization affinities and flexibilities across ARFs.

Introduction

Recent evolutionary and structural studies have elucidated how the family of AUXIN RESPONSE transcription FACTORs (ARFs) originated and has been maintained in the Streptophytes. The last common ancestor of the Streptophytes had two ARFs: a proto-A/B and proto-C ARF, which are still maintained in extant Charophyte algae (reviewed in Carrillo-Carrasco *et al.*, 2023). During Streptophyte evolution, the A/B ARF duplicated and diverged into separate A and B-class ARFs, leading to a last land plant common ancestor with single copies of A, B and C class ARFs. Notably, the extant liverwort *Marchantia polymorpha* still maintains a single copy for each class, and thereby represents the least complex ARF system studied to date. In comparison, the angiosperm *Arabidopsis thaliana* (Arabidopsis) has 5, 15 and 3 copies of A, B and C-ARFs respectively (Bowman *et al.*, 2017; Flores-Sandoval *et al.*, 2015; Mutte *et al.*, 2018; Okushima *et al.*, 2005).

Atomic structures have been experimentally solved for the DNA-binding domains (DBDs) of three ARF proteins: the A-class ARF5 from Arabidopsis and the B-class ARF1 from Arabidopsis and ARF2 from *M. polymorpha* (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Kato *et al.*, 2020). Despite *M. polymorpha* and Arabidopsis diverging approximately 500 million years ago (Harris *et al.*, 2022), DBD structures were remarkably similar. This structural similarity is also reflected by similar capacities and preference for DNA-binding (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Kato *et al.*, 2020). Less is known about the C-ARFs or the charophycean ARFs. Homology modeling indicates these ARFs likely share a similar structure to other ARFs, and both Arabidopsis AtARF10 and a proto-C-ARF from *Chlorokybus atmophyticus* (CaARF) can bind land-plant derived auxin response elements (AuxRE) *in vitro* (Kato *et al.*, 2020; Martin-Arevalillo *et al.*, 2019; Mutte *et al.*, 2018). These studies suggest that the ARF DBD structure and DNA-binding preferences have been maintained in different lineages of the Streptophytes, but is difficult to extrapolate beyond the few ARFs studied.

The protein structures of the ARF DBD revealed key domains that explained the DNA-binding mechanisms of AtARF1, AtARF5 and MpARF2 (see Chapter 2) (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Kato *et al.*, 2020). Briefly, the DBD is composed of a DNA-interacting B3 domain, a dimerization domain (DD) and an ancillary domain (AD) **(Fig. 1A)**. ARFs can heterodimerize and homodimerize through their DD via alpha-helix α6 and can therefore bind bipartite motifs composed of two TGTCNN elements with defined spacing, i.e. direct, everted and inverted repeats (DR, ER and IR). Their spacing preferences are class-specific as A-ARFs can bind a wider variety of IR motifs (with spacing of 5-9 bases: IR5-9), whereas B-ARFs are more restricted and bind only IR7-8 motifs (Boer *et al.*, 2014; Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). The B3 domain is connected to the intertwined DD and AD via a fold of 3 beta sheets of the B3 domain (β4, β9 and β10) and α1 from the DD, which we termed the B3-hinge hereafter. Upon DNA-binding to an IR7 motif, the B3 domain of AtARF1 twists around α1 by 25° to conform to

the DNA (Boer *et al.*, 2014). The dimerization interface does not change conformation during DNA-binding, but ARFs have different affinities for homodimerization, with the A-ARF AtARF5 having a higher affinity than the B-ARF AtARF2 (Fontana *et al.*, 2023). Given that ARFs bind cooperatively to the two binding sites in the target DNA (Fontana *et al.*, 2023; Freire-Rios *et al.*, 2020; Lanctot *et al.*, 2020; Tiwari *et al.*, 2003), it follows that higher dimerization affinity leads to higher DNA-binding affinity. Thus, variation outside of the ARF-DNA interaction interface could also contribute to changes in DNA-binding affinity within the ARF family.

The B3 domain of each individual ARF binds to a TGTCNN AuxRE element (Boer *et al.*, 2014). Both *in vitro* and *in vivo*, ARFs have the highest affinity to TGTCGG motifs, although other motifs are bound and present in auxin-responsive promoters (Boer *et al.*, 2014; Cherenkov *et al.*, 2018; Franco-Zorrilla *et al.*, 2014; Galli *et al.*, 2018; Kato *et al.*, 2020; Lanctot *et al.*, 2020; Lieberman-Lazarovich *et al.*, 2019; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019; Zemlyanskaya *et al.*, 2016). It is well established that land plant A-ARFs and B-ARF preferences to TGTCGG elements is widespread, and recently it was shown that even C-ARFs, whether AtARF10 or the charophycean CaARF, can bind this element (Martin-Arevalillo *et al.*, 2019). Structural and homology modelling studies have shown that this preference is likely conserved in all lineages of the plant kingdom and across the three classes of ARFs, as the DNA-binding residues found by Boer *et al.* (2014) are conserved between several ARFs. This strong conservation suggests there is purifying evolutionary selection against mutations that disrupt binding to TGTCNN elements. However, what remains unclear is: i) how conserved the B3 domain are in general, and in their DNA-binding residues in particular, and ii) if detrimental mutations in the DNA interface occur, whether there are compensating mutations, otherwise known as evolutionary couplings (Clarke, 1995).

Here, we investigated the relationship between conservation of residues and conservation of structure plant kingdom wide. We first used a published high resolution AtARF1 structure to re-evaluate mechanisms by which ARFs interact with the DNA and found that π-interactions and water-mediated contacts are likely important. We look at plant kingdom wide conservation on a per-residue basis to infer that the B3 domain is very strongly conserved whereas the domains involved in dimerization and flexibility are highly variable. We used AlphaFold2 (Jumper *et al.*, 2021), to predict and compare ARF protein structures from Arabidopsis, *M. polymorpha* and several algae and show that the ARF structure is strongly conserved. Finally, we found that the DNA-binding residues have evolutionary couplings with structural and other DNA-binding residues.

Results

High resolution AtARF1 structure reveals an extended DNA-binding interface

The DNA-binding residues in Arabidopsis ARF1 were initially identified in a structure with TGTCTC DNA (PDB: 4ldx; hereafter AtARF1-TC) (Boer *et al.*, 2014). Since then, two more DBD structures have been solved: AtARF1 with TGTCGG DNA (PDB: 6ycq; hereafter AtARF1-GG) and MpARF2 with TGTCGG DNA (PDB: 6sdg; hereafter MpARF2-GG). Since the AtARF1-GG structure has a higher resolution than AtARF1-TC, even individual water molecules could be resolved. The MpARF2-GG structure also allows us to compare conservation of the DNA-binding mechanisms between the B-class ARFs AtARF1 and MpARF2. We therefore reasoned that the availability of AtARF1-GG and MpARF2-GG structures warranted re-evaluation of the DNA-binding interface.

At large, the residues directly contacting the DNA in either the high resolution AtARF1-GG or MpARF2-GG structures are the same as previously identified **(Fig. 1B, 1C)** (Boer *et al.*, 2014). Direct contacts are defined as residues that make direct hydrogen bonds or hydrophobic interactions with the DNA bases, on either the phosphate backbone or within the major groove (ARFs do not interact in the minor groove). We found the previously described AtARF1-R186 makes hydrogen bonds with bases T1 and G2 in both AtARF1 structures. However, while the homologous residue R228 in MpARF2 formed bonds with G2 in one monomer, R228 in the other monomer rotates deeper into the DNA major groove and additionally contacts T3 **(Fig. 1D)**. It is likely that AtARF1-R186 could also make this contact due to the intrinsic rotational flexibility of arginine's. We therefore find that the same residues directly contact the DNA in all three structures, although MpARF2-R228 can potentially adopt alternative rotamers to contact T3, a DNA base critical for DNA-binding by ARFs yet previously not identified to be bound directly by a residue (Boer *et al.*, 2014; Ulmasov *et al.*, 1997a, 1999b).

In the last decade, π-interactions have been identified as additional elements in DNA- and RNA-protein binding (Adhav and Saikrishnan, 2023; Wilson *et al.*, 2014; Wilson *et al.*, 2016). We found three π-interactions likely contributing to DNA recognition by AtARF1, mediated by residues H136, R181 and R186 **(Fig. 1E)**. H136 forms π-π interactions with base G8' (i.e., outside of the TGTCTC element) in a single monomer in AtARF1-TC (not shown), while it forms π-π interactions with C4 in both DBD monomers in the high resolution AtARF1-GG structure. However, the homologous residue MpARF2-H178 has no π-π interactions with the TGTCGG DNA. AtARF1-R181 makes an additional cation-π interaction with base C6' in one monomer in the AtARF1-GG structure, but no such interactions were found for this residue in the other structures. Finally, AtARF1-R186 (MpARF2-R228) makes a cation-π interaction with T1 in one monomer of all three structures. This suggests that π-interactions are an additional element involved in ARF-

DNA binding.

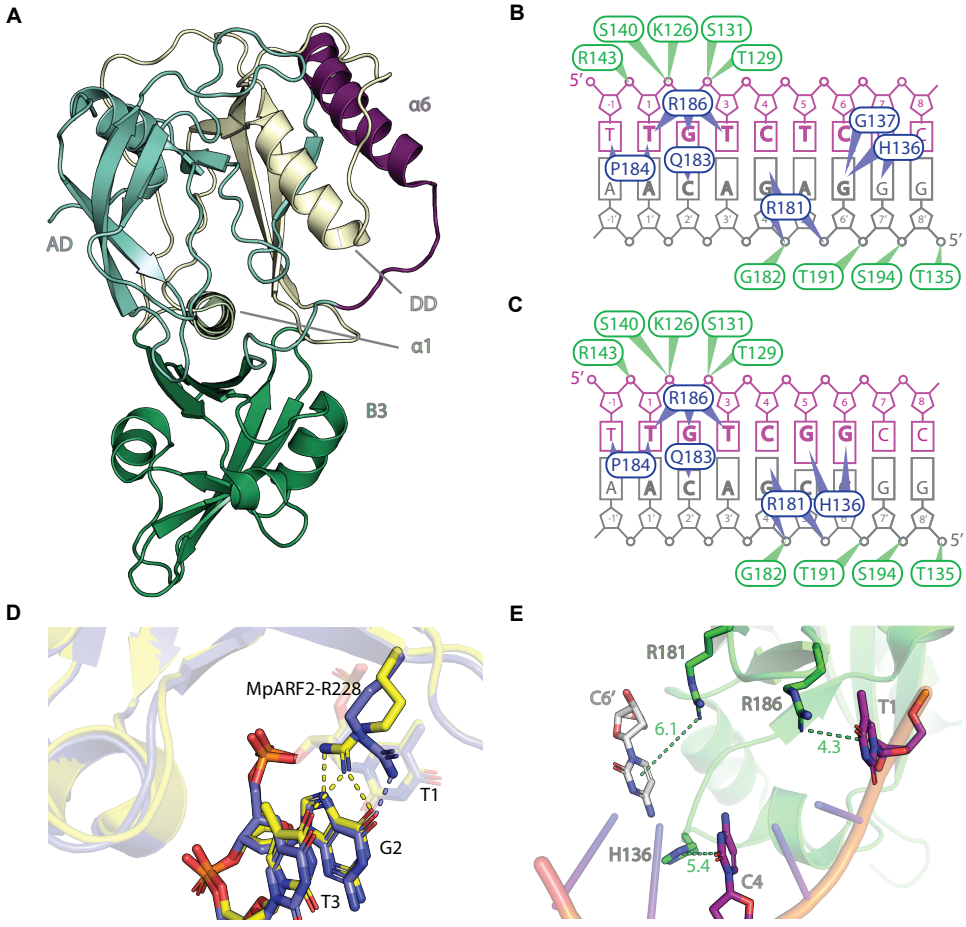


Fig. 1 | Direct, non-covalent interactions in the ARF DNA-binding interface. (A) Cartoon overview of a single AtARF1 DBD with domains of interest highlighted (PDB: 6ycq): DD-B3 hinge (α1, pear), N-terminal dimerization domain (DD, yellow), the dimerization helix (α6, purple), the B3 domain (green) and the ancillary domain (AD, cyan). **(B,C)** Schematic overview of AuxRE recognition by ARF-DBDs. Both figures are composites of contacts in structures 4ldx, 6ycq and 6sdg (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Kato *et al.*, 2020). Residues that make direct hydrogen bonds or hydrophobic interactions to canonical TGTCTC (B) or high affinity TGTCGG (C). Residues bind the nucleobases in the major groove (blue) or the phosphate backbone (green). **(D)** Structural alignment of monomer A and B of MpARF2. Cartoons, carbon atoms from R228 and DNA bases, and hydrogen bonds (dashed lines) are labelled in different colors for each monomer (yellow/marine). **(E)** Pi-interactions (green dashed lines, distance in Å) in structure AtARF1-GG (6ycq) between residues H136, R181 and R186 and DNA bases T1, C4 and C6' respectively. Interacting residues and DNA bases shown as sticks in D and E. Except for D, all figures follow residue numbering of AtARF1.

We additionally found several residues making indirect contacts to DNA through water molecules, both near the protein surface and in the cavity between the ARF and DNA major groove. As water molecules near the protein surface are constantly moving under normal physiological temperatures, we only considered residues that use a single water molecule as a 'bridge' to make contacts to the DNA. For instance, T191 forms direct hydrogen bonds with the phosphate group of C6' while both T190 and R177 require water molecules to bind to C5' **(Fig. 2A)**. In total, we found six residues that bind the phosphate backbone via a water bridge (A130, D132, R144, R177, R185 and T190) and two residues (T129 and S194) that bind both directly **(Fig. 1B, 1C)** and indirectly. We identified 15 water molecules occupy the cavity between the B3 domain and the major groove (not shown). Four of the residues that make direct contacts to the nucleobases additionally use water molecules to bridge to other nucleobases (H136, R181, Q183 and R186). For instance, H136 in AtARF1-GG forms a π-π interaction with base C4 **(Fig. 1E)**, forms hydrogen bonds with G6, and a single water molecule connects H136 and G7'

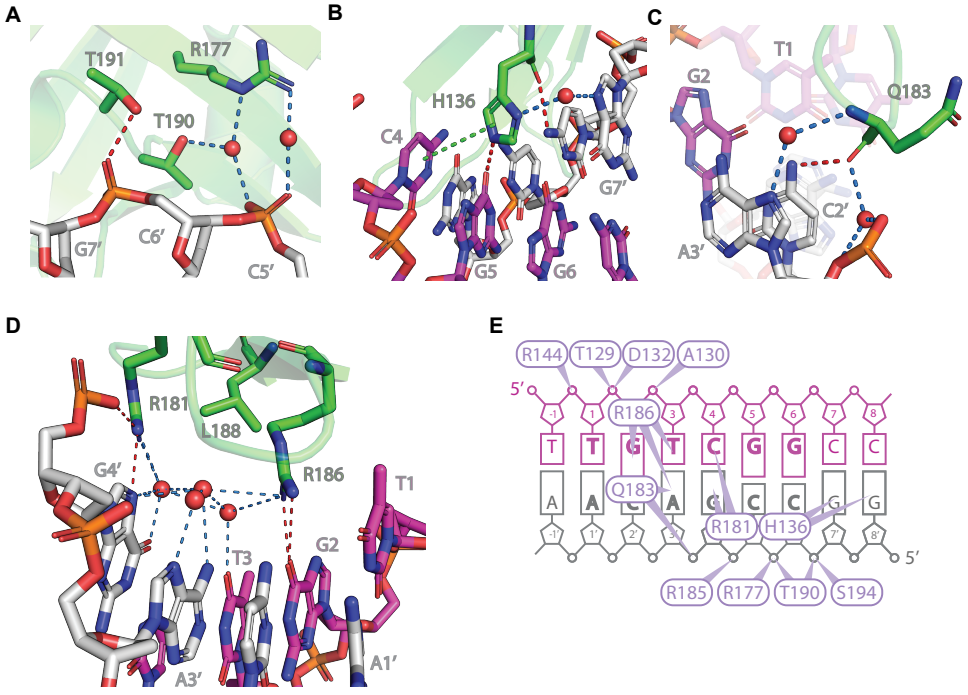


Fig. 2 | Water-mediated hydrogen bonds in the ARF DNA-binding interface. (A-D) Closeups of water-mediated interactions between AtARF1 and TGTCGG AuxRE (PDB: 6ycq): (A) R177, T190 and T191; (B) H136; (C) Q183; and (D) R181, R186 and L188. Carbon atoms are colored according to AtARF1 (green), DNA leading strand (magenta) and DNA lagging strand (silver). Water molecules are shown as red spheres. Interactions (dashed lines) indicate π-interactions (green), hydrogen bonds between AtARF1 residues and DNA directly (red) or water-mediated hydrogen bonds (marine). **(E)** Schematic overview AtARF1 residues (purple circles) forming water-mediated hydrogen bonds to DNA bases.

(Fig. 2B). Moreover, the carboxyl group of Q183 forms direct hydrogen bonds to C2' and additionally binds indirectly to A3' at the phosphate backbone and the nucleobase via two water molecules **(Fig. 2C)**. Finally, we identified that R181, R186 and L188 trap four water molecules that form an extensive hydrogen bond network with bases T3:A' and G4' **(Fig. 2D)**. Together, these results suggests that waters play an integral part of ARF DNA recognition **(Fig. 2E)**.

The ARF B3 domain is strongly conserved in the plant kingdom

With the DNA-binding residues identified we investigated to what extent were these residues conserved across the plant kingdom. We analyzed the conservation of AtARF1-DBD at the residue level and mapped this directly onto the ARF1-GG protein structure **(Fig. 3A, S1, S2)**, using the ConSurf webserver (Ashkenazy *et al.*, 2016). The entire DBD has a conservation score of 5.47±2.92 (mean conservation score ± standard deviation), of which the B3 domain was highly conserved (7.52±1.95). The DD in comparison was more variable (4.65±2.84). The most variable residues in the DD appear to be solvent-facing whereas conserved residues seem to be mostly buried **(Fig. S1)**. Indeed, solvent accessible residues (accessibility >30%) have lower conservation scores (3.31±2.36) than buried residues (accessibility ≤30%; 5.73±2.75). Moreover, we found deeply buried residues (accessibility ≤10%) to have slightly higher conservation (6.21±2.70).

When we investigated important domains in the DBD, we found that the B3-DNA interface showed strong conservation for the DNA-binding residues **(Fig. 3B, Table 1)**, suggesting that TGTCNN recognition mechanism is strongly conserved in the ARF family. Similarly to the DD, the buried residues appeared slightly more conserved (8.28±1.20) than the accessible residues (6.75±2.25), implying evolutionary pressure on functional and structural residues within the B3 domains. We find that B3-hinge interface between the B3 and DD is only partially conserved **(Fig. 3C)**. The interface consists primarily of short hydrophobic residues (A, I, L, and V) akin to a velcro strap. While the C-terminal part of α1 is strongly conserved (AtARF1: L19-G29; right-side in **Fig. 3C**), including the corresponding interacting residues of β4, β9 and β10 from the B3 domain, the N-terminal part is highly variable. Since α1 is 19 residues long in AtARF1 (P11-G29, ARF1-GG) and only 10 residues long in AtARF5 (N52-G62, ARF5-TC) (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020), it is possible that the observed variation in AtARF1 residues P11-L19 corresponds with varying lengths of α1 and thereby differences in flexibility of the interface. The dimerization interface shows the highest variability in AtARF1 **(Fig. 3D)**, including the dimerization residues (conservation score: 4.65±2.21).

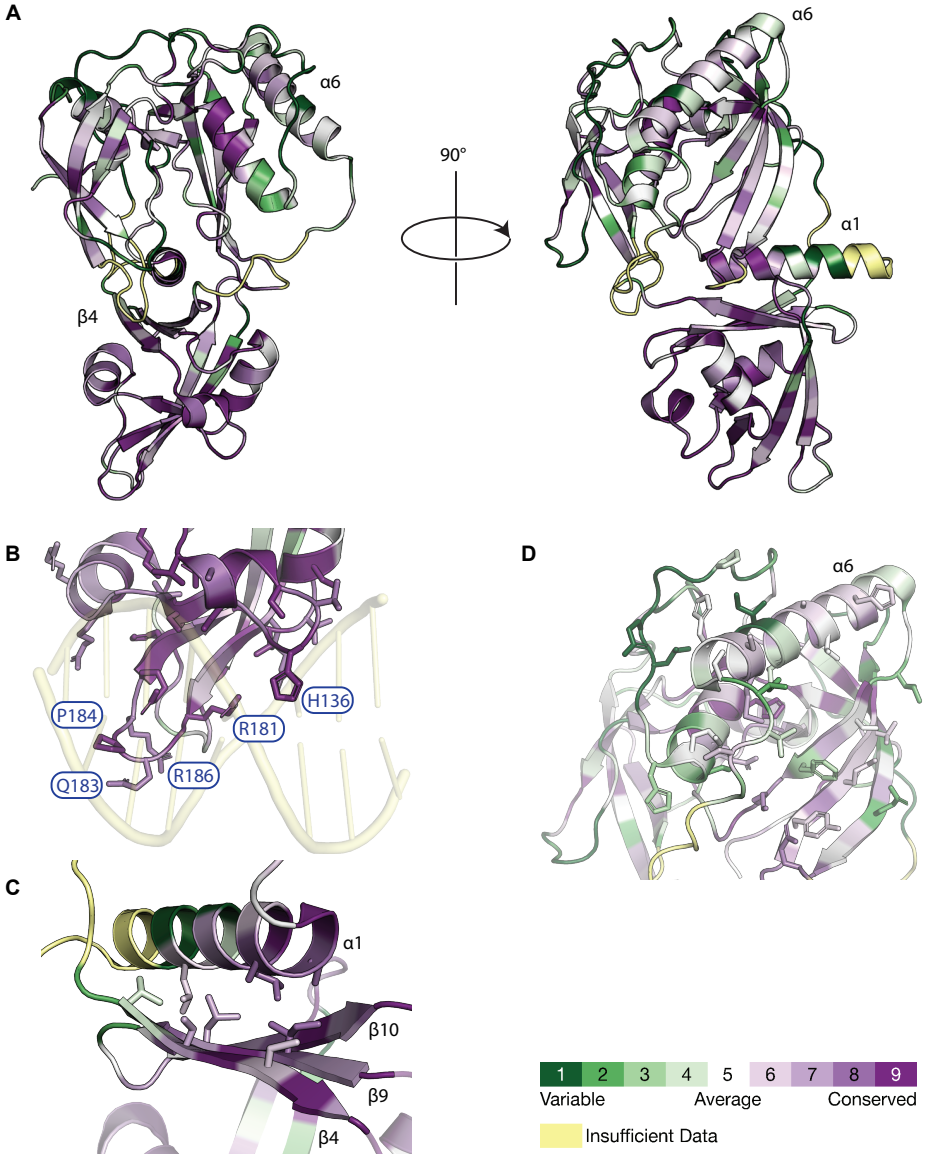


Fig. 3 | Conservation of individual residues in ARF DBD. ConSurf analysis (Ashkenazy *et al.*, 2016) mapped onto the AtARF1-GG structure (6ycq). **(A)** Cartoon representation of AtARF1-DBD from two angles. The dimerization helix (α6) and the B3-Dimerization Domain (DD) hinge domain (α1 and β4) are indicated. **(B)** Closeup of the B3 domain, with DNA-interacting residues shown as sticks, and major groove interacting residues numbered. **(C)** Closeup of the B3-DD hinge. Hydrophobic residues at the interface shown as sticks. **(D)** Closeup of the dimerization interface. Residues involved in homodimerization are shown as sticks (Boer *et al.*, 2014). (A-D) ConSurf conservation scores per residue range from 1-9 (green to purple), higher score is more conserved. Yellow, insufficient data.

ARF DBD structure is conserved in the plant kingdom

Given that we find that structural residues in the DBD appear to be strongly conserved among ARFs, we evaluated how conserved the ARF structure is in the plant kingdom by using AlphaFold2 (AF) to predict and compare structures of different ARFs (Jumper *et al.*, 2021). We initially explored ARF structures in *M. polymorpha*, since its set of 3 ARFs (1 each for each class) allows full comparison of structures within a species (Bowman *et al.*, 2017; Flores-Sandoval *et al.*, 2015; Kato *et al.*, 2015). Although the DBDs only share 60%

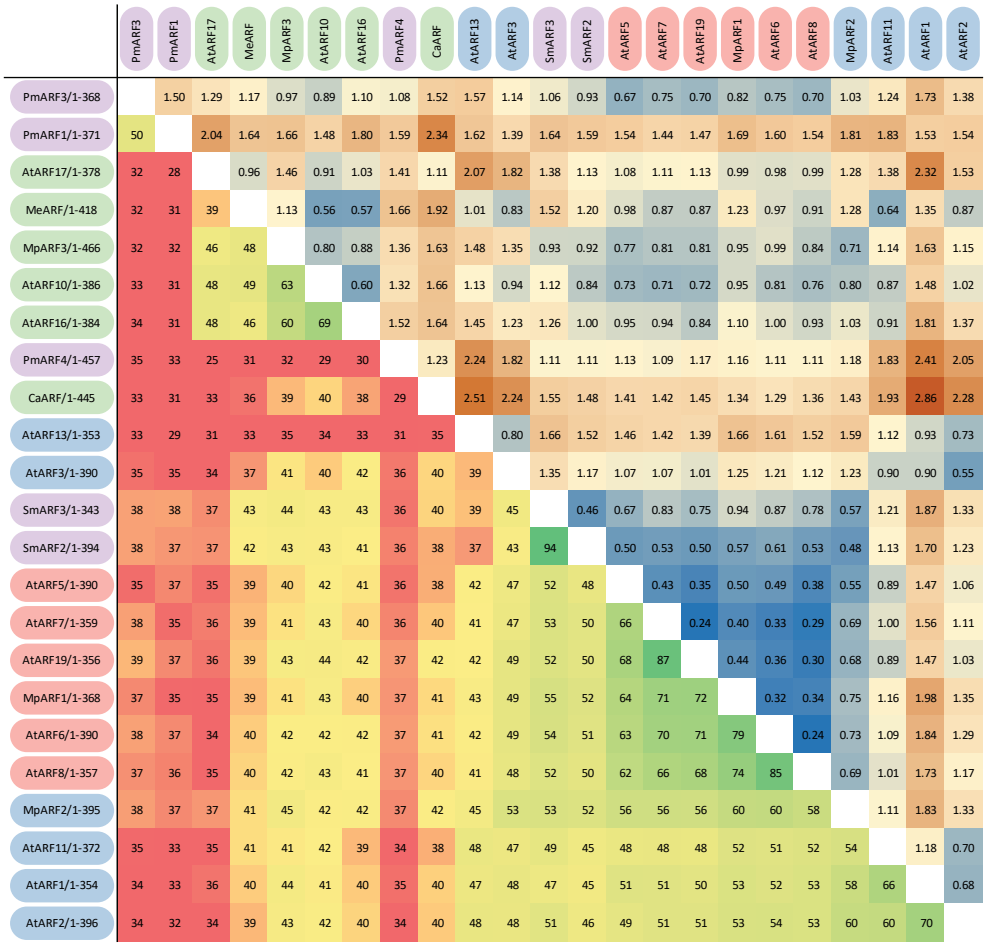


Fig. 4 | Crosstable of percentage identity and RMSD of the DBDs of ARFs used in this study. ARFs were clustered by similarity according to Clustal Omega. Numbers on the lower triangle shows percentage (%) identity (red to green). Upper triangle shows RMSD of each AlphaFold2 predicted structure with the other predicted structures (blue to orange). Column and row titles show ARF/residue numbers used for the comparison. Background color behind each ARF indicates their class: A/B (purple), A (red), B (blue) and C (green). At: Arabidopsis thaliana, Ca: Chlorokybus atmophyticus, Me: Mesotaenium endlicherianum, Mp: Marchantia polymorpha, Pm: Penium margaritaceaum, Sm: Spirogloea muscicola.

(MpARF1 to MpARF2), 41% (MpARF1 to MpARF3), and 45% (MpARF2 to MpARF3) protein sequence identity, their structures have only 0.75, 0.95 and 0.71 Å root mean square deviation (RMSD) respectively **(Fig. 4)**. Thus, ARF DBD structures from the different classes have the same overall structure, although we do note that MpARF3 has a long, disordered loop connecting the dimerization helix to the B3 domain (AtARF1 α6 to β4; **Fig. 5A**). These similarities extend to the DNA-binding residues and their positioning **(Fig. 5B)**. To examine whether this similarity is retained in species with expanded ARF families, we compared the protein sequence identity and AF predictions of ARFs from representative ARFs of each subclade of Arabidopsis **(Fig. 4)**. For the land plant ARFs, we found that even with a percentage identity as low as 31% (AtARF13 versus AtARF17), the largest RMSD found is only 2.32 Å (AtARF1 versus AtARF17). Class A (MpARF1 and AtARF5/6/7/8/19) are particularly similar in both their sequence identity (62-87%) and structure (max 0.5 Å RMSD). These results suggest that the ARF-DBD structure has been conserved in land plants both between classes of ARFs, as well as in expanded families of each class, despite low sequence identities.

To analyze if ARF-DBD structural conservation in land plants extends to extant charophycean ARFs, we additionally predicted structures of DBDs from different lineages of Charophyte algae **(Fig. 4, 6)**; from the Zygnematophyceae, A/B-class DBDs from *Penium margaritaceaum* (PmARF1/2/3/4) and *Spirogloea muscicola* (SmARF1/2/3), and C-class DBDs from *Mesotaenium endlicherianum* (MeARF) and the Chlorokybophyceae *Chlorokybus atmophyticus* (CaARF). PmARF2 and SmARF1 had incomplete sequence annotations, leading to structures lacking domains in the AF predictions and were therefore excluded from comparison. The algal DBDs have only low percentage sequence identity with land plant DBDs **(Fig. 4)**, with MeARF below 50% and CaARF, PmARF and

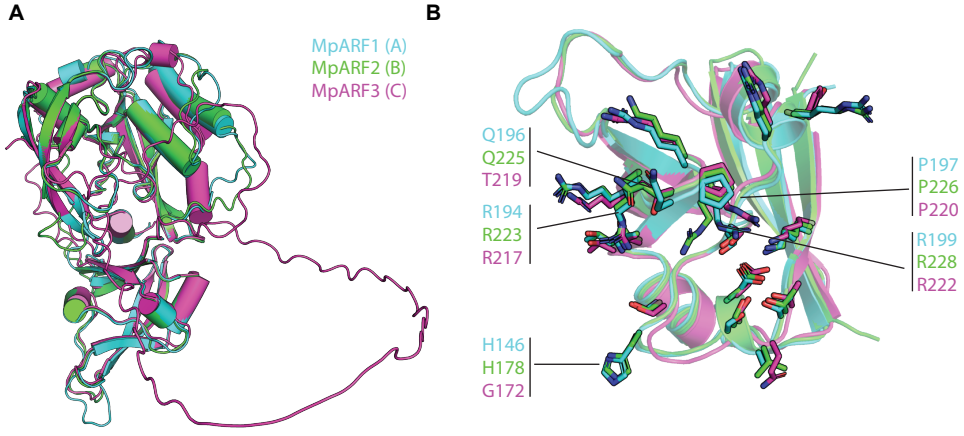


Fig. 5 | AlphaFold2 predicted M. polymorpha ARF DBDs. (A) Structural alignment of AlphaFold2 (Jumper *et al.*, 2021; Varadi *et al.*, 2021) predicted DBDs of MpARF1 (A-class, cyan), MpARF2 (B-class, green) and MpARF3 (C-class, magenta). **(B)** Closeup of DNA-binding interfaces. AtARF1-homologous residues in MpARFs shown as sticks. Major groove binding residues are highlighted.

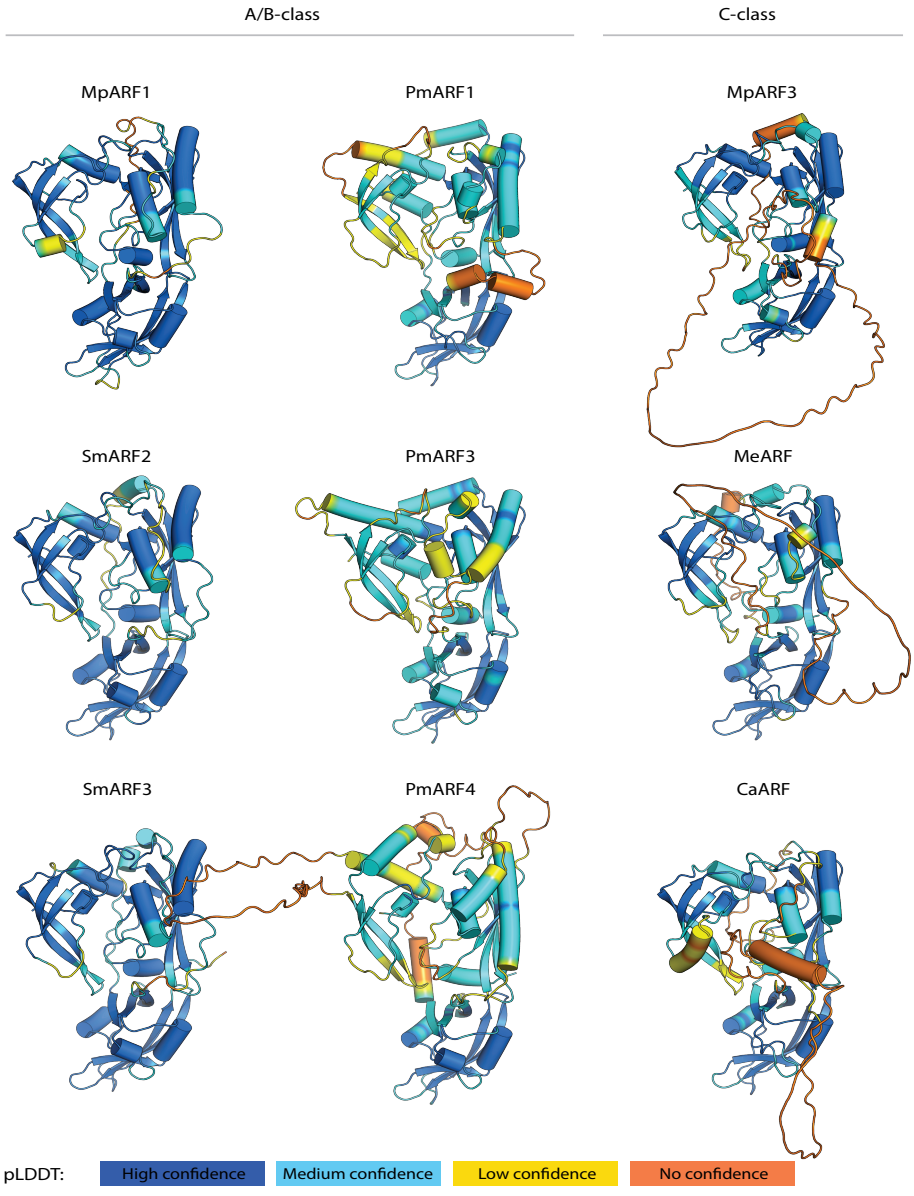


Fig. 6 | AlphaFold2 predicted algal ARF DBDs. ARFs are classified as A/B-class and C-class. All predicted ARF DBDs are viewed from the same angle, with the dimerization helix (α6 in AtARF1) on the top right, the ancillary domain (AD) on the top left and the B3 domain at the bottom. Colors represent AlphaFold2 per-residue accuracy (pLDDT). MpARF1 and MpARF3 were included as reference structures and are the same as in Fig. 5. Ca: *Chlorokybus atmophyticus*, Me: *Mesotaenium endlicherianum*, Mp: *Marchantia polymorpha*, Pm: *Penium margaritaceaum*, Sm: *Spirogloea muscicola*.

SmARFs only below 40%. Structurally however, the charophycean ARF-DBDs resemble land plant ARFs with a maximum RMSD of 2.86 Å (CaARF versus AtARF1) **(Fig. 4, 6)**. CaARF-DBD has between 29-41% sequence identity to all other ARF-DBDs, although structurally it maintains the ARF-DBD fold and the C-class-specific extended loop. On sequence identity, MeARF-DBD is more similar to land plant C-class (39-49%) than any other ARF-DBDs (31-42%), including CaARF (36%) **(Fig. 4)**. This is reflected in the structures, having the C-class loop and having an RMSD of only 0.56 and 0.57 with DBDs from AtARF10 and AtARF16 respectively. The PmARF-DBDs have the lowest sequence identity overall to land plant ARFs among all tested DBDs (25-39%) and have the highest similarity to land plant A-class ARFs (36-39%). Even between the PmARFs, sequence identities are low (33-50%). Overall, the PmARFs resemble the land-plant ARF-DBD structures **(Fig. 4, 6)** with PmARF3 closely resembling A-class ARFs (RMSD 0.67-0.82 Å). However, we do note that the DD of the PmARFs show altered topologies compared to land plant DBDs, including different angles of the dimerization helix for PmARF3 and PmARF4 **(Fig. 6)**. Since the confidence value of these predictions is low, we additionally predicted their structures with ESMFold (Lin *et al.*, 2023). The AF and ESMFold predictions of MpARF3 and PmARF4 were exact matches (not shown), suggesting there is cryptic information in the sequences that trigger both machine learning algorithms to predict alternative topologies. Finally, SmARF2 and SmARF3 share 94% sequence identity, implying a recent duplication event and share both the the highest sequence identity (48-55%) and structure (RMSD 0.94-0.46 Å) with land plant A-class DBDs **(Fig. 4)**. Together, these predictions suggest that despite millions of years of independent evolution, ARF-DBDs maintained similar structures since their origin in the streptophyte last common ancestor.

We then investigated whether the charophycean ARF-DBDs also conserved the DNA-binding residues. Since the B3 domains, including the DNA-binding residues, were strongly conserved **(Fig. 3B, 5B)**, we calculated the sequence similarity for just the B3 domains **(Fig. S3)**. Indeed, the B3 domains showed a higher percentage similarity than the entire DBDs **(Fig. 4)**, and we additionally observed AF having high confidence in the B3 domains **(Fig. 6)**. Here, we found that the DNA-binding residues are conserved at sequence level and structurally in the charophycean ARF-DBDs **(Fig. 7)**. This all reinforces that the B3 domain structure and DNA-binding residues are strongly conserved in the plant kingdom, implying strong evolutionary pressure to maintain DNA-binding.

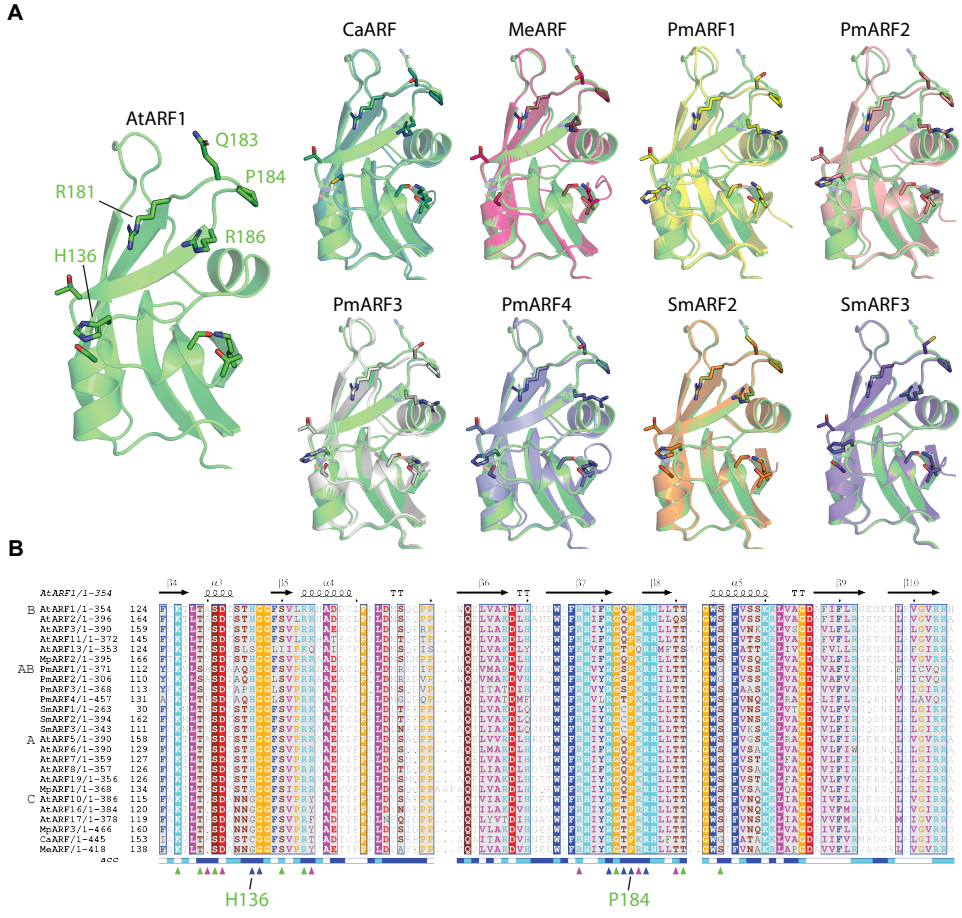


Fig. 7 | Closeup of the B3 domain and the DNA-binding residues of A/B-ARFs and C-ARFs. (A) AlphaFold2 predicted structures for ARF B3 domains. Protein structure of AtARF1-GG (6ycq) is shown on the left (green) with direct DNA-binding residues as sticks, and major groove binding residues named. Predictions are aligned to the AtARF1 structure and homologous DNA-binding residues are shown as sticks. C-ARFs: CaARF and MeARF. A/B-ARFs: PmARFs and SmARFs. (B) Alignment of B3 domains of ARFs. The secondary structure of AtARF1 (6ycq) is depicted above the alignment. ACC below the alignment shows the solvent accessibility (white to blue, from low to high accessibility) based on the AtARF1 structure (6ycq). Arrows indicate DNA-binding residues: major groove binders (blue), phosphate backbone binders (green), and water-bridged residues (purple). At: *Arabidopsis thaliana*, Ca: *Chlorokybus atmophyticus*, Me: *Mesotaenium endlicherianum*, Mp: *Marchantia polymorpha*, Pm: *Penium margaritaceaum*, Sm: *Spirogloea muscicola*.

A subset of DNA-binding residues is co-evolving

Missense mutations can alter or abolish protein functionality, unless co-occurring compensatory mutations allow proteins to regain their original function, leading to co-evolution of two residues, or evolutionary coupling. Although the DNA-binding residues of the ARFs are highly conserved, we wondered whether any of these residues had couplings. We used the EVCoupling webserver (Hopf *et al.*, 2019), using both MpARF1 (A-class) and MpARF2 (B-class) as 'seeds' to find shared and class-specific couplings. In total, we found 42 and 47 couplings for MpARF1 and AtARF1 respectively **(Fig. S4)**. For DNA-binding residues of AtARF1 specifically, 17 couplings involve DNA-binding residues, and 7 couplings are found for the major groove-binding residues.

For the major groove-binding residues, we found AtARF1-P184 coupled with S140, T190, W193 and G205, of which the former two are involved in binding the phosphate backbone and the latter two are likely structural **(Fig. 8A)**. We found the same couplings for MpARF1-P194, although it has no coupling with S150 (homologous to AtARF1-S140). Prolines are important structural residues in proteins due to their rigidity and it is therefore possible that these couplings are required to compensate for a change in structure upon mutation of P184. A similar situation was found with AtARF1-G137, S140, R181 and L188 which are coupled and all are related to both DNA-binding and trapping the waters between the B3 domain and the DNA **(Fig. 8B)**. Finally,

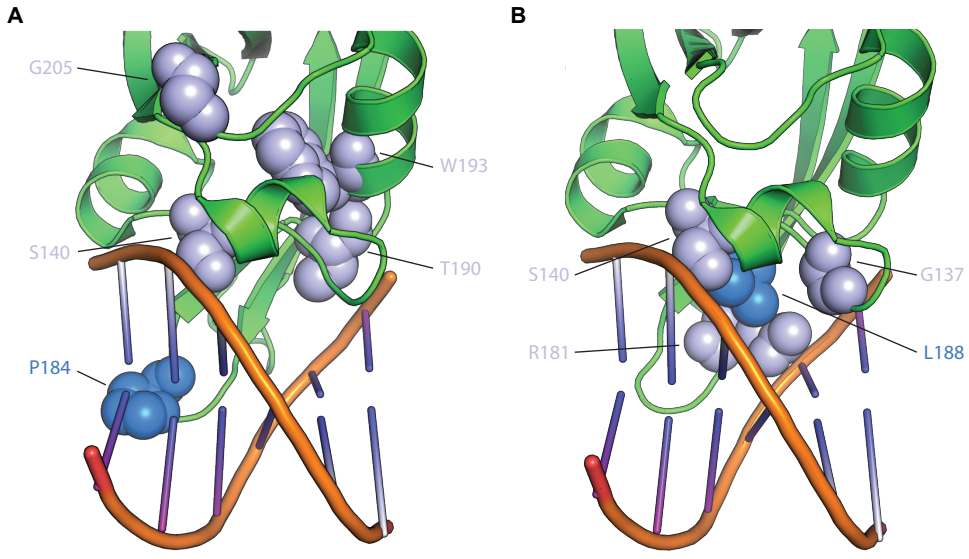


Fig. 8 | Evolutionary couplings in A-class and B-class ARFs mapped onto AtARF1. Evolutionary coupling determined with the EVCoupling webserver (Hopf *et al.*, 2019) using AtARF1 and MpARF1 as 'seeds'. **(A-B)** Cartoon representation of B3 (green) and AuxRE (orange) interface. Coupled residues indicated as spheres. (A) L188 (blue) is coupled with G137, S140 and R181 (grey). (B) P184 (blue) is coupled with S140, T190, W193 and G205 (grey).

while AtARF1-R186 has no couplings, MpARF1-R199 is coupled with the hydrophobic core residue W206 (AtARF1-W193). This coupling is unclear but could imply an A-ARF specific coupling. Of note, we did not find couplings for AtARF1-H136 (MpARF1-H146), despite that we observed variation in the underlying multiple sequence alignment (not shown) and the residue showing variation in Arabidopsis ARFs (Boer *et al.*, 2014). Finally, although all major groove binders except H136 have couplings, we did not observe any couplings between the major groove binders.

Table 1 | Summary of DNA-binding residues, their conservation and any evolutionary couplings.

Residue					
AtARF1	**MpARF1**	**MpARF2**	**Binds***	**Consurf****	**EVC*****
K126	K136	K168	pl	9	
T129	T139	T171	pl	9	
A130	A140	V172	water	8	D147, K306
S131	S141	S173	pl	9	
D132	D142	D174	water	9	
T135	T145	T177	pl	8	
H136	H146	H178	mjgr	9	
G137	G147	G179	mjgr	9	**R181**, L188
S140	S150	S182	pl	9	**P184**, L188, **S194**
R143	R153	R185	pl	8	**R185**
R144	R154	R186	water	7	C125
R177	R190	R219	water	8	P159, **S194**
R181	R194	R223	mjgr	9	**G137**, L188
G182	G195	G224	pl	9	
Q183	Q196	Q225	mjgr	8	T133
P184	P197	P226	mjgr	9	**S140**, **T190**, W193, G205
R185	R198	R227	water	8	**R143**, I179
R186	R199	R228	mjgr	9	
L188	L201	L230	****	8	**G137**, **S140**, **R181**
T190	T203	T232	water	9	**P184**
T191	T204	T233	pl	8	
S194	S207	S236	pl	9	**S140**, **R177**

* Residue interacts with the phosphate backbone (pl), the major groove (mjgr) or binds the DNA via water-mediated hydrogen bonding (water).

** Higher Consurf Scores indicate higher conservation (range 1 to 9).

*** Couplings in Bold are to other DNA-binding residues. Numbers refer to AtARF1 residues.

**** L188 is indirectly involved (Fig. 2D).

Discussion

The ARFs are present in each lineage of the Streptophytes. Based on a range of phylogeny, protein structures, homology modelling, *in vitro* DNA-binding studies and *in vivo* studies, a picture emerged that the structure and DNA-binding preferences of ARFs are likely deeply conserved (Boer *et al.*, 2014; Fontana *et al.*, 2023; Freire-Rios *et al.*, 2020; Galli *et al.*, 2018; Kato *et al.*, 2020; Martin-Arevalillo *et al.*, 2019; Mutte *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). However, questions remained about the relationship between structural conservation and conservation at the residue level. In this chapter, we used a combination of detailed analysis of available structures, sequence and (predicted) structural comparisons to explore the ARF-DNA interface. While multiple structures have been reported (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Kato *et al.*, 2020), the description of protein-DNA contacts was largely based on a single structure with lower resolution than the more recent AtARF1-GG structure. We therefore reanalyzed the DNA-binding interface of AtARF1 and MpARF2 with high affinity TGTCGG DNA and revealed the interface is more extensive than was previously identified. Previous studies highlighted the importance of the core TGTC and mutations on base T3 would disrupt recognition by AtARF1, despite no residue making base contacts in the original AtARF1-TC structure (Boer *et al.*, 2014; Ulmasov *et al.*, 1997a, 1999b). Here, we identified that base T3 is contacted in the MpARF2 structure by a deeper rotation of residue R228 into the major groove, while in AtARF1-GG we identified a water-mediated hydrogen bond network to T3:A3'. These results suggest that the arginine and water molecules are flexible and can adopt multiple states. Water molecules play an integral part in DNA binding by nucleosomes and DNA recognition by the human estrogen receptor (ER), the latter for which molecular dynamics showed three water molecules were crucial for sequence specificity, as the protein was essentially immobilizing the three water molecules (Davey *et al.*, 2002; Kosztin *et al.*, 1997; van Dijk *et al.*, 2013). Molecular dynamics could therefore be used to fully understand the B3-DNA interface to see: i) if the water molecules, R181 and R186 are immobilized or flexible, and in the latter case, ii) if mutations in the core TGTC nucleotides create a steric hindrance and detach the ARFs from the DNA and whether water molecules are involved in that process.

Although the ARF DBDs are structurally similar, they are functionally distinct, as neither the MpARF2 nor MpARF3 DBD can completely functionally replace the ARF1 DBD in mutant complementation assays (Kato *et al.*, 2020). In this work, we find that the B3-hinge and the dimerization domains show residue variation **(Fig. 3)**. Since A-ARFs and B-ARFs show different preferences for bipartite motifs (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Galli *et al.*, 2018; Kato *et al.*, 2020; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019) and we find the B3-hinge interface is only partially conserved, it is likely that the variable residues create different flexibilities in this domain. This could be experimentally tested by performing α1 helix swaps between AtARF1 and AtARF5 and

measure DNA binding preferences. We found that the dimerization domain is highly variable in the plant kingdom. It was shown that AtARF1 lacking the PB1 domain and full length AtARF4 can heterodimerize on IR7 motifs *in vitro* (Ulmasov *et al.*, 1999b) and that homodimerization affinities differ between AtARF2 and AtARF5 (Fontana *et al.*, 2023). The lack of conservation in this domain may therefore suggest that multiple DD-mediated interaction interfaces could be formed. However, the biological relevance of ARF heterodimerization is entirely unknown, and this is an important direction for future investigation.

We found that the B3 domain and the DNA-binding residues are strongly conserved in all Streptophyte lineages **(Fig. 3B, 5B, 7, S4)**, which corroborates previous studies that suggest ARFs are universally capable of binding TGTCNN elements (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Galli *et al.*, 2018; Kato *et al.*, 2020; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). Thus, ARF binding to TGTCNN elements is likely an ancient trait facing strong evolutionary pressure to be preserved, even in largely expanded families like Arabidopsis. A testable prediction from this observation is that B3 domains should be interchangeable between ARFs. Furthermore, we observed evolutionary couplings for the DNA-binding residues, including the major groove interacting residues **(Fig. 8, Table 1)**. This finding implies coevolution in the B3 to maintain binding to TGTCNN elements, but would require functional validation, for instance by making single and double substitutions.

In summary, our analyses indicate that the ARF B3 domain and the DNA-binding residues are strongly conserved in the plant kingdom, suggesting that binding to TGTCNN elements is an ancestral property of ARFs. We have also found that residues involved in dimerization and intrinsic flexibility of the B3 domain are very variable in the plant kingdom, implying differences in spacing preferences and homodimerization affinities are widespread. Furthermore, the DNA-binding residues are evolutionary coupled, suggesting that binding to TGTCNN elements can be achieved through multiple modes. These results offer an extended structural framework for exploring the mechanisms that dictate ARF-DNA binding specificity and its biological relevance.

Material & Methods

DNA-binding residues identification

DNA-binding residues were identified on the previously solved structures of AtARF1 DBD (6ycq) (Freire-Rios *et al.*, 2020). Direct interactions were identified in PyMOL Molecular Graphics System (version 2.5.5, Schrödinger, LLC) using the *distance* command between the ARF and the DNA molecules, in mode 2 to detect polar contacts, and with a cutoff of 3.2 Å for hydrogen bonds. For indirect interactions (water to water, water to DNA and water to ARF), *solvent* was used as an additional selection in the *distance* command. Hydrophobic interactions between thymine and hydrophobic residues (i.e. P184) were visually checked via VdW radii. Finally, π-interactions were also determined with the *distance* command, in mode 5 with a cutoff of 6.5 Å.

Structure conservation analyses

Conservation score values for ARF DBD residues were calculated using the ConSurf webserver (Ashkenazy *et al.*, 2016). AtARF1 (6ycq, chain A) was used as query for the HMMER search (1 iteration), 0.0001 E-value cutoff, UNIREF-90 database). Next, 150 sequences that sample the list of homologues were selected for further analyses after filtering % identity (between 35-95%). These sequences were aligned with MAFFT (L-INS-i) and the evolutionary conservation scores per residue (ranging from 1-9 for highly variable [1], intermediate conservation [5] or highly conserved [9]) calculated under a Bayesian method with the best substitution model. Average conservation scores and standard deviations for specific domains were calculated by subsetting each domain and extracting conservation scores. To distinguish buried and accessibly residues we calculated relative solvent accessibility per residue in PyMOL with the command *get_sasa_relative*. Buried residues were defined as 0-30% relative accessibility, and accessible residues as 31-100% relative accessibility. Deeply buried residues were defined as 0-10% relative accessibility.

Protein structure prediction

Structures of ARFs were predicted using AlphaFold2 (AF) and ESMFold (Jumper *et al.*, 2021; Lin *et al.*, 2023) with default settings. Available full-length ARF sequences were used for structure prediction, and middle region and PB1 manually trimmed to cover the DBD. RMSD was calculated for the trimmed structures with the *extra_fit* command, using *align* set to 10 cycles to remove outliers.

Performance of AF for *de novo* ARF structure prediction was manually evaluated. First, AtARF1 structure was predicted and compared with the existing apo- and DNA-bound AtARF1 (PDB: 4ldv, 4ldx) structures. Root mean square deviation (RMSD) similarity confirmed AF prediction (RSMD of 1.53 or 1.34 Å for apo and DNA-bound respectively).

Second, AF was examined for *de novo* prediction of AF-unknown ARF structures. For this, MpARF2 (58% and 56% shared sequence identity with AtARF1 and AtARF5 respectively, **(Fig. 4)** was used due to having a solved structure (PDB: 6sdg) that was absent in the AF training dataset. AF prediction with default settings matched well with the solved structure (RMSD of 0.86 Å). In a default prediction, AF searches the PDB for structures homologous to the query to aid refinement of the prediction, i.e., it uses MpARF2 (6sdg) in a default prediction. We therefore limited this search to exclude MpARF2 (PBD pre-2019) and MpARF2, AtARF1 and AtARF5 (PDB pre-2013) and confirmed that even without refinement, AF is capable of predicting MpARF2 *de novo* (RMSD of 0.86 and 1.67 Å for pre-2019 and pre-2013 respectively).

Evolutionary coupling analyses

Evolutionary coupling for DBD residues was analyzed using the EVCoupling webserver (Hopf *et al.*, 2019). The DBD sequences of MpARF1 (M1-P368) and AtARF1 (M1-P354) were independently used as queries or seeds. Settings were modified to include sequences with at least 85% length as the queries to exclude unrelated B3 domains from the alignments. Identical sequences were removed to avoid anecdotal coupling amplification. Sequence searches were ran against the UniRef100 database and retrieved sequences with a bitscore of 0.7 or lower were kept leaving 545 and 540 sequences for MpARF1 an AtARF1 respectively. Alignments were manually confirmed to contain only ARF sequences and cover three classes (A,B and C). Evolutionary coupling search was ran using a 75% probability threshold.

Supplementals

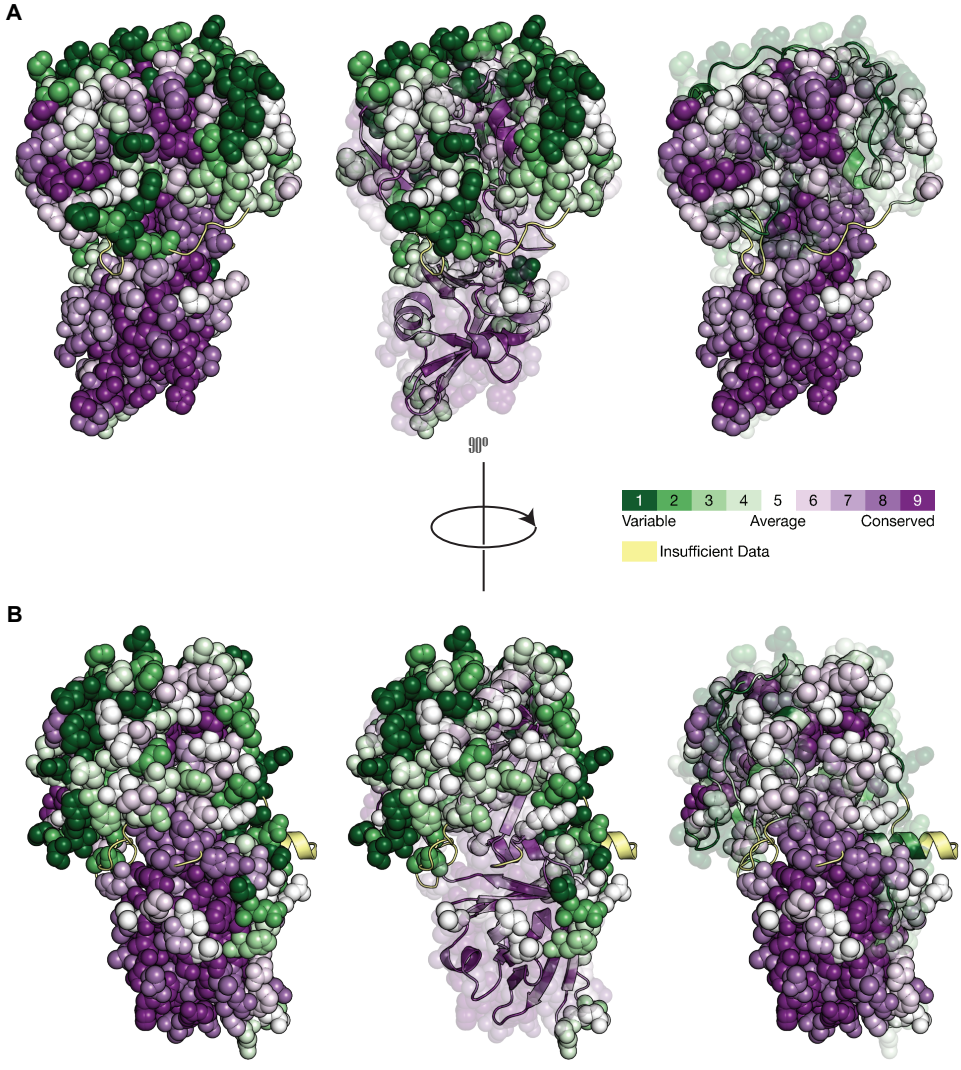


Fig. S1 | Sphere representation of conserved residues mapped on the AtARF1 structure. (A,B) Sphere representation overlayed on cartoon representation of AtARF1 structure (PDB:6ycq) (Freire-Rios *et al.*, 2020) from same angles as in Fig. 3. AtARF1 was used in a default ConSurf analysis (Ashkenazy *et al.*, 2016). ConSurf gives conservation scores from 1-9, green to purple, higher is more conserved. Yellow represents insufficient information. Left: all residues shown as spheres. Middle: conserved residues (ConSurf ≤ 5) transparent. Right: variable residues (ConSurf ≥ 5) transparent.

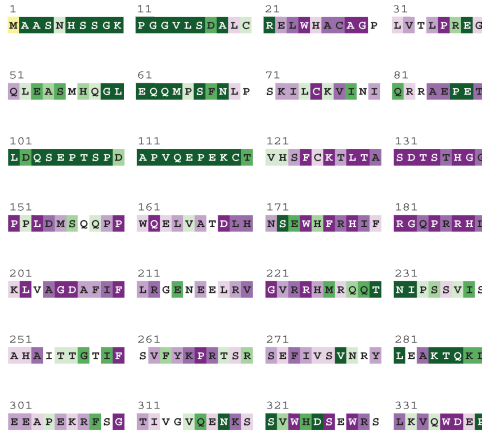


Fig. S2 | Conservation of individual residues mapped on the AtARF1 protein sequence. AtARF1 was used in a default ConSurf analysis (Ashkenazy *et al.*, 2016). ConSurf gives conservation scores from 1-9, green to purple, higher is more conserved. Yellow represents insufficient information.

	PmARF3	PmARF1	CaARF	AtARF17	MeARF	MpARF3	AtARF10	AtARF16	AtARF13	PmARF4	SmARF3	SmARF2	AtARF11	MpARF2	AtARF1	AtARF2	AtARF3	MpARF1	AtARF6	AtARF8	AtARF5	AtARF7	AtARF19
PmARF3/113-217		71	54	50	50	50	52	52	50	59	55	55	55	59	56	54	58	55	55	54	55	57	58
PmARF1/112-215	71		51	47	51	52	48	49	39	55	55	55	52	56	53	51	54	53	55	54	56	54	54
CaARF/153-257	54	51		50	59	62	58	59	47	50	60	60	54	59	60	57	61	59	59	59	54	54	58
AtARF17/119-223	50	47	50		63	66	65	65	41	49	54	55	51	53	54	50	54	53	51	52	55	54	52
MeARF/138-242	50	51	59	63		84	79	75	47	53	58	60	59	61	59	59	65	63	60	60	62	59	59
MpARF3/160-264	50	52	62	66	84		87	84	46	52	58	60	57	62	62	61	65	59	58	57	59	58	59
AtARF10/115-219	52	48	58	65	79	87		90	49	50	59	59	56	61	57	57	61	61	59	60	62	63	63
AtARF16/120-224	52	49	59	65	75	84	90		47	49	60	60	55	61	57	57	61	60	59	60	60	60	62
AtARF13/124-230	50	39	47	41	47	46	49	47		46	48	49	60	56	55	60	52	53	49	48	48	49	50
PmARF4/131-235	59	55	50	49	53	52	50	49	46		58	61	53	61	58	57	56	54	56	55	57	56	55
SmARF3/111-215	55	55	60	54	58	58	59	60	48	58		96	62	69	64	65	66	68	69	68	70	70	70
SmARF2/162-266	55	55	60	55	60	60	59	60	49	61	96		64	70	65	66	67	69	69	68	70	70	70
AtARF11/145-249	55	52	54	51	59	57	56	55	60	53	62	64		73	76	78	70	63	61	62	64	58	59
MpARF2/166-270	59	56	59	53	61	62	61	61	56	61	69	70	73		80	83	74	72	73	72	73	73	73
AtARF1/124-228	56	53	60	54	59	62	57	57	55	58	64	65	76	80		89	72	71	70	70	70	67	69
AtARF2/164-268	54	51	57	50	59	61	57	57	60	57	65	66	78	83	89		70	70	70	70	71	69	70
AtARF3/159-263	58	54	61	54	65	65	61	61	52	56	66	67	70	74	72	70		75	73	70	70	68	69
MpARF1/134-241	55	53	59	53	63	59	61	60	53	54	68	69	63	72	71	70	75		90	84	82	85	85
AtARF6/129-233	55	55	59	51	60	58	59	59	49	56	69	69	61	73	70	70	73	90		90	80	83	82
AtARF8/126-230	54	54	59	52	60	57	60	60	48	55	68	68	62	72	70	70	70	84	90		77	78	80
AtARF5/158-262	55	56	54	55	62	59	62	60	48	57	70	70	64	73	70	71	70	82	80	77		83	85
AtARF7/127-231	57	54	54	54	59	58	63	60	49	56	70	70	58	73	67	69	68	85	83	78	83		91
AtARF19/126-230	58	54	58	52	59	59	63	62	50	55	70	70	59	73	69	70	69	85	82	80	85	91	

Fig. S3 | Percentage identity of the B3 of ARFs used in this study. ARFs were clustered by similarity according to Clustal Omega. Numbers indicate percentage (%) sequence identity (amino acids) with a color scale from red to green. Column and row titles are ARF/residue numbers used for the comparison. Background color behind each ARF indicates their class: A/B (purple), A (red), B (blue) and C (green). At: *Arabidopsis thaliana*, Ca: *Chlorokybus atmophyticus*, Me: *Mesotaenium endlicherianum*, Mp: *Marchantia polymorpha*, Pm: *Penium margaritaceaum*, Sm: *Spirogloea muscicola*.

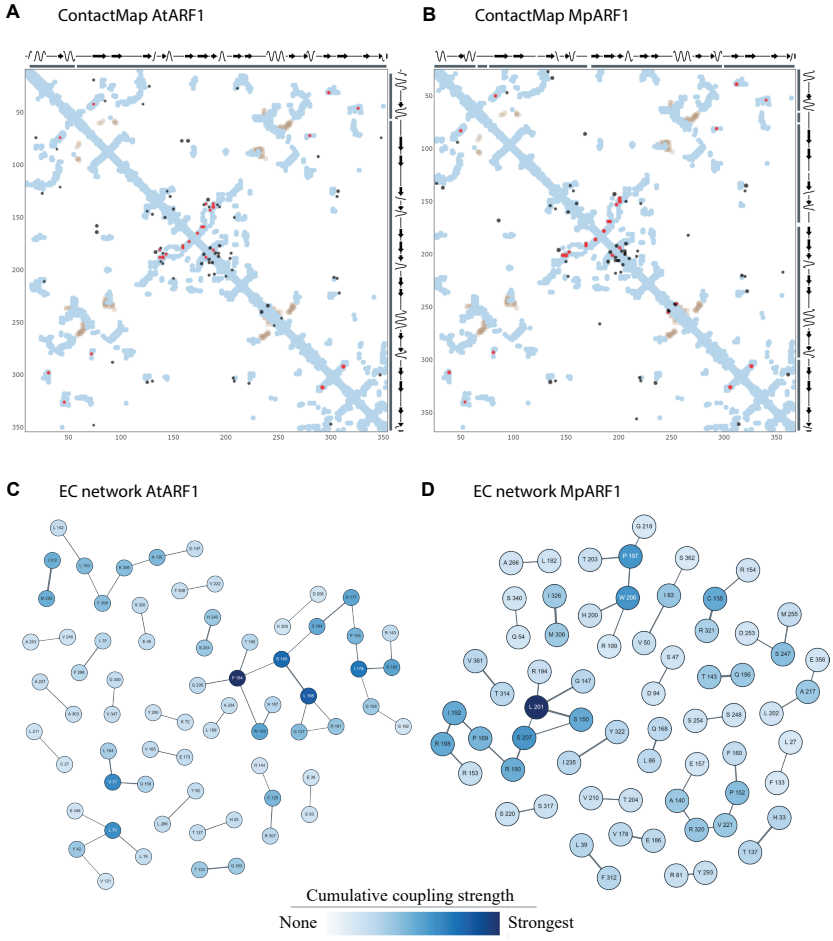


Fig. S4 | Evolutionary couplings (EC) in A-class and B-class ARF DBDs. (A, B) Contactmaps of AtARF1 (A) and MpARF1 (B). Dots indicate whether residues: are in contact in the structure (lightblue); facilitate homodimerization (brown); have EC and are within 5 Å of each other (red); or have EC and are more than 5 Å from each other (black). Dotsize indicates probability of a real EC, between 75-100% probability. **(C, D)** EC network maps of AtARF1 (C) and MpARF1 (D). Nodes are residues that have EC with others and color represents cumulative coupling strength, with the strongest residues (e.g. L188 in C) having the most EC with other residues. Edges represent EC between residues. The wider the edge, the higher the probability (between 75-100%) for a real EC.

Chapter

4

DNA-binding preferences of the *Marchantia polymorpha* ARF1 transcription factor

Juriaan Rienstra[1], Victor Levistky[2,3], Sumanth Mutte[1], Shubhajit Das[1], Pingtao Ding[4], Victoria Mironova[2,5], Jorge Hernández-García[1] and Dolf Weijers[1]

[1] Laboratory of Biochemistry, Wageningen University and Research, Wageningen, the Netherlands

[2] Department of Systems Biology, Institute of Cytology and Genetics, Novosibirsk, Russia

[3] Department of Natural Science, Novosibirsk State University, Novosibirks, Russia

4 Institute of Biology, Leiden University, Leiden, the Netherlands

5 Department of Plant Systems Physiology, Radboud Institute for Biological and Environmental Sciences (RIBES), Radboud University, Nijmegen, Netherlands

Abstract

AUXIN RESPONSE FACTORs (ARFs) are the focal point in the nuclear auxin signalling pathway, selecting the output of auxin response through specific DNA binding. In *Marchantia polymorpha*, with the least complex NAP studied to date, the A-class MpARF1 and B-class MpARF2 compete for shared DNA sites. However, the target genes of MpARF1 and 2 are not known, and neither are their intrinsic DNA-binding preferences or the extent of similarity with angiosperm ARFs. Here, we use a combination of transcriptomics, heterologous and in vitro protein-DNA interaction assays to explore the binding specificity and target genes of MpARF1. We find that, like angiosperm ARFs, MpARF1 prefers TGTCGG elements and can bind a direct repeat with 5 bases spacing (DR5) in a bacterial one-hybrid assay. However, from a combination of motif-enrichment analyses in *M. polymorpha* auxin-treated transcriptomes and of MpARF1 genome-wide binding preferences we found MpARF1 to likely bind DR[-2] and DR2 motifs. Structural modelling suggests mechanisms through which MpARF1 homodimers may bind these motifs, with neither model relying on the canonical dimerization helix. We speculate that a last common A-class ARF ancestor likely had features of both MpARF1 and angiosperm ARFs, that diverged between the two different clades.

Introduction

As the DNA-binding transcription factors in the pathway, AUXIN RESPONSE FACTORs (ARFs) are central to the nuclear auxin response pathway (NAP). *Marchantia polymorpha* has the least complex NAP studied to date, with one copy each of the auxin receptor Mp*TIR1*, the transcriptional repressor Mp*IAA*, and each of the A, B and C ARF classes (Mp*ARF1,* Mp*ARF2,* and Mp*ARF3* respectively) (Bowman *et al.*, 2017; Flores-Sandoval *et al.*, 2015; Kato *et al.*, 2015). In the last decade, several studies have used this low level of complexity to elucidate the core principles of the NAP (for a more detailed description see Chapter 2). Under low cellular auxin conditions, MpARF1 acts as a repressor by interacting with MpIAA, which in turn recruits the transcriptional co-repressor MpTOPLESS (MpTPL). At higher auxin conditions, MpIAA is degraded through interaction with MpTIR1, liberating MpARF1 and turning it into a transcriptional activator. MpARF2 recruits MpTPL directly, thereby acting as a repressor regardless of auxin concentrations (Flores-Sandoval *et al.*, 2015; Kato *et al.*, 2015; Kato *et al.*, 2017; Kato *et al.*, 2020; Suzuki *et al.*, 2023). Overexpression of either Mp*ARF1* or Mp*ARF2* led to aberrant phenotypes, whereas simultaneous overexpression reverted to a wild type phenotype (Kato *et al.*, 2020). It was therefore hypothesized that these two ARFs are competing on the same DNA sites, an observation that was previously made for the more complex NAP of the moss *Physcomitrella patens* (Lavy *et al.*, 2016). However, in a domain swap experiment, replacing the DNA-binding domain (DBD) of Mp*ARF1* with the DBD of Mp*ARF2* failed to fully complement the Mp*arf1-4* null mutant (Kato *et al.*, 2020), suggesting their DNA-binding profiles only partly overlap. Moreover, maize (*Zea mays*) A-ARFs and B-ARFs seem to bind in the vicinity to each other, rather than directly compete on the same DNA elements (Galli *et al.*, 2018). These results indicate that the MpARF1 and MpARF2 DNA-binding profiles likely have some degree of overlap, although the full extent is unknown. Understanding the DNA-binding of *M. polymorpha* ARFs would reveal the simplest DNA-binding competition model and allow for comparative analyses with angiosperm ARFs.

DNA-binding preferences have been studied extensively for angiosperm A- and B-class ARFs, while preferences of bryophyte ARFs are largely unknown. As described in Chapter 2, ARFs are modular proteins composed of three domains: a DNA-binding domain (DBD), a middle region (MR) and Phox and Bem1 domain (PB1). The DBD contains a DNA-interacting B3 domain that binds an auxin responsive element (AuxRE) in the DNA. For angiosperm A- and B-class ARFs, the highest affinity AuxRE is a TGTCGG element, although both can bind a range of TGTCNN elements with lower affinities (Boer *et al.*, 2014). This is reflected in the Arabidopsis genome, as auxin-responsive genes contain a variety of TGTCNN elements in their promoters (Freire-Rios *et al.*, 2020; Lieberman-Lazarovich *et al.*, 2019; Zemlyanskaya *et al.*, 2016). Both MpARF1 and MpARF2 can bind TGTCGG elements and share conserved DNA-interacting residues with angiosperm ARFs (Chapter 3) (Kato *et al.*, 2020; Mutte *et al.*, 2018), although it is unknown whether this is

their highest affinity element. In addition to direct DNA-binding, ARFs can homodimerize via their DBDs, facilitating cooperative binding to bipartite AuxRE motifs (Boer *et al.*, 2014; Fontana *et al.*, 2023; Freire-Rios *et al.*, 2020; Kato *et al.*, 2020). These two AuxRE's can be in three different conformations – direct, inverted and everted repeats (DR, IR and ER respectively) – with specific base spacing (e.g. an inverted repeat with seven bases between the AuxRE is an IR7). See Chapter 2 for more information. Based on in vitro binding profiles, angiosperm A-class ARFs seem to bind a wider variety of motifs than B-class ARFs (Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). For instance, while both *Arabidopsis thaliana* (Arabidopsis) B-class AtARF2 and A-class AtARF5 can bind IR7, AtARF5 can additionally bind bipartite motifs like DR5, IR7/8 and ER13 (O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). The DR5 tandem repeat motifs have been the basis for a widely used auxin response reporter in Arabidopsis due to its high auxin sensitivity, a sensitivity that could be explained by DR5 preferentially being targeted by the auxin-responsive A-class ARFs (Freire-Rios *et al.*, 2020; Ulmasov *et al.*, 1997b). For the *M. polymorpha* ARFs, both MpARF1 and MpARF2 have been shown to bind IR7 motifs *in vitro* (Kato *et al.*, 2020), however their spacing preferences are unknown. Revealing the MpARF DNA-binding preferences could therefore reveal the extent to which land plant ARFs follow an evolutionary conserved DNA-binding preference.

In this chapter, we investigated the DNA-binding preferences of MpARF1. We confirm that MpARF1 can bind to DR5 motifs in a heterologous system, although we find that MpARF1 has different preferences for bipartite motifs than angiosperm A-class ARFs. We further confirm that some of these different bipartite motifs are enriched in *M. polymorpha* auxin-induced genes. Finally, we corroborate that MpARF1 is likely the sole ARF through which auxin signalling is relayed.

Results

MpARF1 binds DR5 motifs in heterologous systems

Since a previous study found that MpARF1 was able to bind IR7 (Kato *et al.*, 2020), we investigated here whether MpARF1 was capable of binding other known ARF binding motifs using a bacterial one-hybrid (B1H) approach. In our setup, the 'bait' plasmids contained the MpARF1 DBD and the 'prey' vectors contained either a *3xDR5* or a *2xIR8/ER13* motif. We included the DBD of Arabidopsis ARF5 (also known as MONOPTEROS) (Boer *et al.*, 2014; Fontana *et al.*, 2023; Freire-Rios *et al.*, 2020; Guilfoyle *et al.*, 1998; Hardtke and Berleth, 1998; Ulmasov *et al.*, 1999a, b) as a control for a well-studied flowering plant A-class ARF. We observed that both AtARF5 and MpARF1 were able to activate on the *3xDR5* prey plasmid and not on the *2xIR8/ER13* prey **(Fig. 1A)**. To investigate whether the growth of *E.coli* was caused by binding of AtARF5 or MpARF1 to the *3xDR5* prey, we tested the setup with mutant in AtARF5 in which a key DNA-binding residue was mutated, and that had previously been shown to impair specific DNA binding (AtARF5-R215A; REF). As expected, strains carrying this mutant ARF were unable to grow **(Fig. 1B)**. We next engineered the corresponding mutation (R194A) into MpARF1. Similarly, this mutation prevented growth and hence binding of MpARF1 to its target site. This finding demonstrates that MpARF1-DBD can bind a *DR5* motif, and in addition, suggests that it uses a mechanism similar to AtARF5 for binding to DNA.

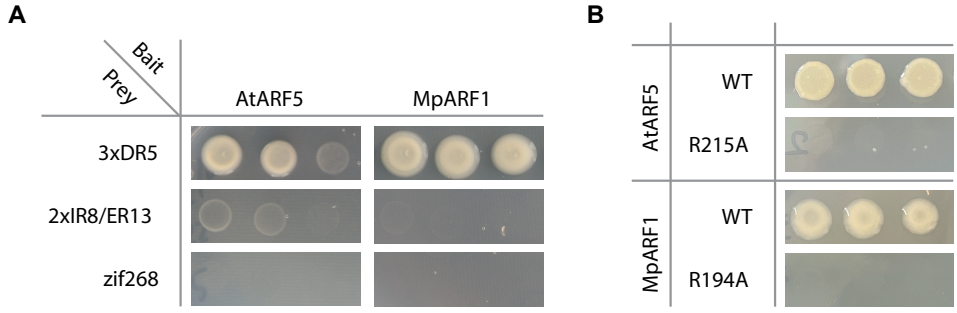


Fig. 1 | Bacterial 1 Hybrid assays with AtARF5 and MpARF1. (A) Bait vectors with AtARF5-DBD and MpARF1-DBD versus prey vectors with 3xDR5, 2xIR8/IR13 and zif268 (negative controls) promoter motifs. **(B)** Bait vectors with AtARF5-DBD and MpARF1-DBD and their respective non-DNA-binding mutants versus the 3xDR5 prey. In all images, from left to right are 10-fold dilutions (10^0, 10^{-1} and 10^{-2} respectively). Note that figure B is the same as Chapter 5, Fig. B1H. Assays were repeated twice with identical results.

Identification of AuxRE's in promoters of auxin-responsive genes

Given that *M. polymorpha* has a single A-class ARF, the current model for auxin-dependent gene regulation would predict that all auxin-regulated gene expression depends on MpARF1. We tested this prediction by compared the transcriptomes of wild-type (male) Tak-1 and the Mp*arf1-4* loss of function mutant (Kato *et al.*, 2017) after exogenous auxin application, by RNA-seq (see details of RNAseq dataset in Chapter 5). Briefly, Tak-1 and Mp*arf1-4* gemmae were treated with the auxin biosynthesis inhibitors L-kynurenine and Yucasin, followed by treatment with indole-3-acetic acid (IAA). We found 112 genes to be differentially expressed in response to IAA in the wild-type **(Fig. 2A)**, a number that is consistent with similar reported datasets **(Fig. 3 A)** (Kuhn *et al.*, 2022; Mutte *et al.*, 2018). Strikingly, under these conditions, we found a complete lack of differential gene expression in the Mp*arf1-4* mutant **(Fig. 2A)**. Indeed, the Log_2 fold-changes of these 112 genes between Tak-1 and Mp*arf1-4* are not correlated (R_2 = 0.032), supporting the

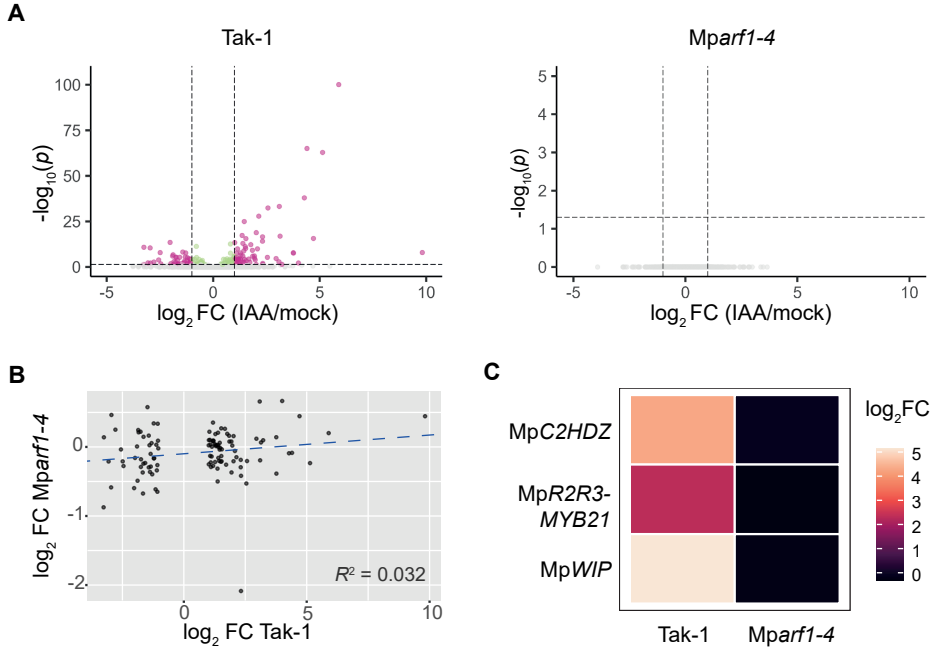


Fig. 2 | Comparison of auxin-dependent gene expression in Tak-1 and Mp*arf1-4* from dataset 1. (A) Volcano plots of auxin-dependent differentially expressed genes in Tak-1 (left) and Mp*arf1-4* (right). Magenta, genes that show significant log_2 fold changes (≤ -1 and ≥ 1) and statistically significant differential expression (multiple testing corrected *p*-value ≤ 0.05); Green, genes without significant log_2 fold changes but with statistically significant differential expression; Grey, rest. **(B)** Scatterplot of log_2 fold changes of auxin-dependent differentially expressed genes in Tak-1 (magenta circles in A) versus corresponding values in Mp*arf1-4*. Blue line, Pearson correlation. R^2 is the Pearson coefficient of determination. **(C)** Log_2 fold changes of auxin-induced genes in Tak-1 and Mp*arf1-4*.

lack of IAA-mediated transcriptional response in the mutant **(Fig. 2B)**. Previous studies found Mp*C2HDZ*, Mp*R2R3* and Mp*WIP* to be auxin-induced in *M. polymorpha* (Flores-Sandoval *et al.*, 2018a; Kato *et al.*, 2017; Mutte *et al.*, 2018; Suzuki *et al.*, 2023). In our experiment, we found that while these genes are strongly induced in the wild-type, they are unresponsive in the mutant **(Fig. 2C)**. Together, these results indicate that MpARF1 is indeed the sole transcription factor mediating IAA transcriptional response in *M. polymorpha*.

Since auxin-regulated transcription is mediated by MpARF1, we reasoned promoters of auxin-responsive genes should be enriched in MpARF1 binding motifs. Therefore, we analyzed five different transcriptome datasets of auxin-treated wild-type plants for AuxRE motif enrichment. In addition to the dataset described above (dataset 1; Chapter 5), we generated two datasets of dormant and non-dormant gemmae treated with IAA (dataset 2 and 3), as gemmae undergo large transcriptional reprogramming after gemmae germination that may gate auxin-responsiveness (Wang *et al.*, 2023). We further included two published datasets covering 10-day-old gemmalings treated with the synthetic auxin 2,4-D or IAA (dataset 4 and 5) (Kuhn *et al.*, 2022; Mutte *et al.*, 2018). Across datasets, we found between 57 and 152 genes to be auxin-responsive **(Fig. 3A)**, which represents 0.25 to 0.65% of the *M. polymorpha* genes (23,399 annotated genes, genome build MpTak_v6.1). Additionally, the datasets have little overlap **(Fig. 3B)**, with only three auxin-responsive genes overlapping between all datasets, no genes overlapping among the gemmaling datasets (dataset 4 and 5), and six genes overlapping among the gemmae datasets (datasets 1, 2 and 3). This low level of overlap suggests that additional factors (environment, developmental timing, experimental setups) may strongly influence auxin-responsiveness in *M. polymorpha*. We next analyzed whether the promoters of the auxin-responsive genes were enriched for known auxin-responsive elements and motifs. We were unable to find specific TGTCNN elements (e.g., TGTCGG) to be enriched, likely due to the small number of auxin-responsive genes. We then performed an enrichment analysis for bipartite TGTCNN motifs in the promoters and 5' untranslated regions (5'UTR) upstream of the transcription start site (TSS) of auxin-responsive genes, for each dataset separately **(Fig. S1)**, followed by a meta-analysis **(Fig. 3C)**. From the meta-analysis, it followed that auxin-induced genes were significantly enriched in motifs DR[-2] (TGTCTGTCNN), DR[-1] (TGTCNTGTCNN) and DR2. In addition, we found a weak enrichment (below the significance threshold; meta-*p*-value < 0.05) for motifs DR4/8/13, IR7 and ER4. No motifs were enriched in the meta-analysis in the auxin-repressed genes, although some motifs (e.g., DR0, ER1/6 and IR12/15) were weakly enriched in individual datasets **(Fig. S1)**. Thus, despite a limited number of auxin-responsive genes in *M. polymorpha*, we identified potential AuxRE's that may act as MpARF1 binding sites.

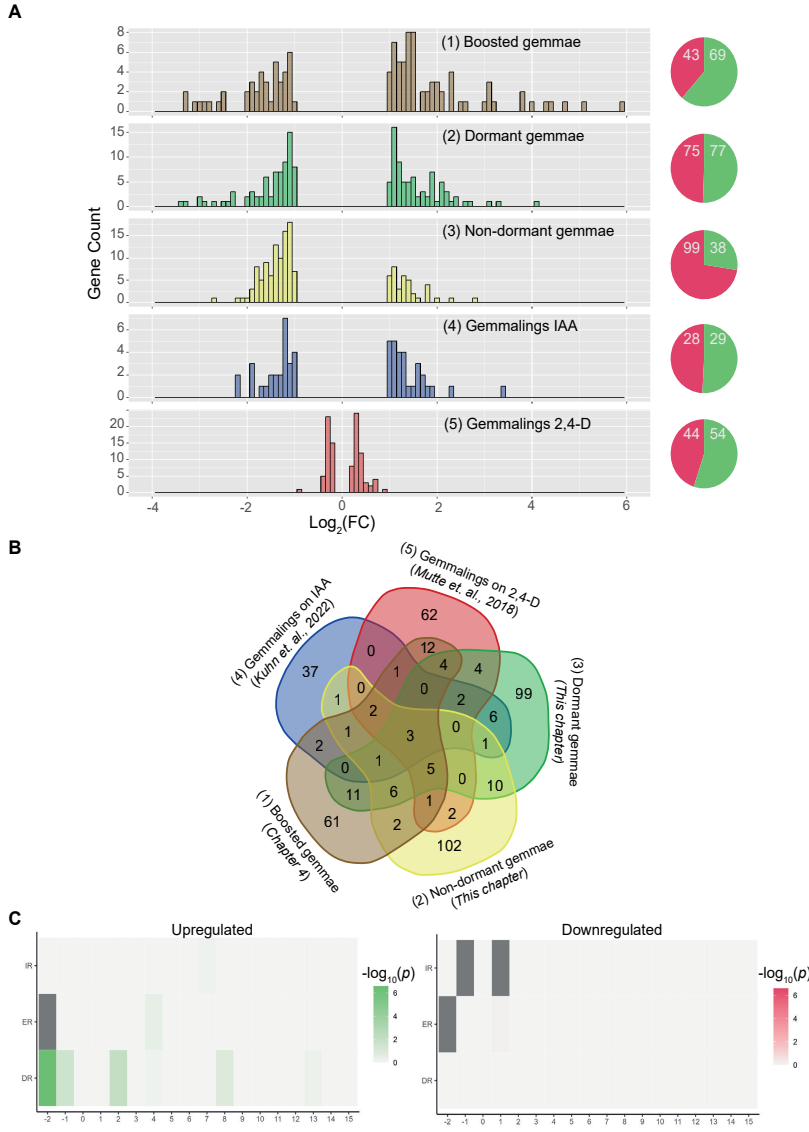


Fig. 3 | *M. polymorpha* auxin-dependent gene regulation. (A) Histograms of $\log_2$ fold change among auxin-dependent differentially expressed genes. Pie charts indicate the number of auxin-induced (green) and auxin-repressed (magenta) genes in each dataset. For dataset 1, 2, 4 and 5, genes were filtered for $p_{adj} \leq 0.05$ and $\log_2$ of ≤ -1 and ≥ 1. Dataset 3, due to low $\log_2$ values, was filtered for $p_{adj} \leq 0.01$. **(B)** Venn diagram indicating the overlap of auxin-dependent differentially expressed genes in the different datasets. **(C)** Enrichment of DR, ER and IR motifs with indicated spacing lengths in the upstream (3,500 bp to transcription start site) of the auxin-responsive genes. Color saturation represents the significance for each bipartite motif in $-\log_{10}$(meta *p*-value). Auxin-induced (green) and auxin-repressed (magenta) genes. Dark grey represents bipartite motifs not found in any promoters.

MpARF1 DNA-binding preferences

To test whether the enriched motifs identified in the RNA-seq meta-analysis were also the preferred binding sites for MpARF1, we performed DNA Affinity Purification and sequencing (DAP-seq), a method previously used on angiosperm ARFs (Chapter 2) (Galli *et al.*, 2018; O'Malley *et al.*, 2016). Due to technical challenges, we could only generate a single replicate. We used the top 5000 peaks for the subsequent analysis. Most of the peaks were found in the upstream regions of genes (29.5%) and coding sequences (CDS; 30.6%) **(Fig. 4A)**. Indeed, the peaks show enrichment primarily 2000 bp downstream of the TSS **(Fig. 4B)**. This finding is in contrast to DAP-seq results with angiosperm ARFs, which primarily bind upstream of the TSS (Galli *et al.*, 2018; O'Malley *et al.*, 2016). The highest enriched element in the MpARF1 peaks was the high affinity AuxRE TGTCGG **(Fig. 4C)**, occurring in 66% of the peaks, which coincides with the preferences of ARFs from Arabidopsis and maize (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014; Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). We then looked for enrichment of bipartite TGTCNN motifs and found that the highest enriched motifs were DR[-2], DR2, DR6 and ER1 **(Fig. 4D)**. The DR motifs are primarily found enriched around the TSS, while the ER1 motif is spread evenly **(Fig. 4E)**.

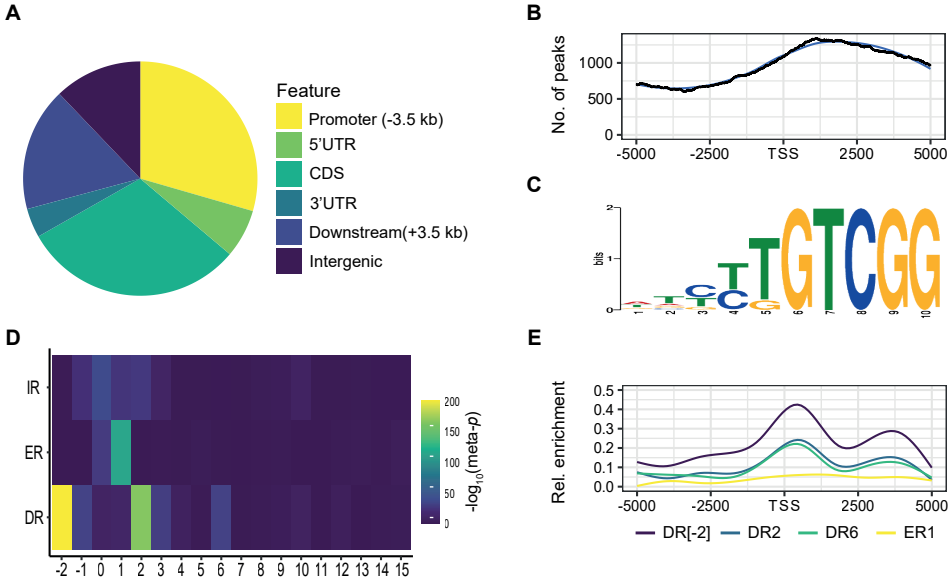


Fig. 4 | MpARF1 binding preferences. The top 5000 peaks from a DAP-seq analysis with MpARF1 were used for the following analyses. **(A)** Percentage of peaks found in different genomic features. Promoters are defined as 3,500 bp upstream of the transcription start site (TSS), downstream as 3,500 bp downstream of the coding sequence (CDS) stop codon. **(B)** Distribution of peak summit around the TSS of raw data (black) and locally weighted Loess smoothing (blue). **(C)** Position Weight Matrix of top enriched motif. **(D)** Enrichment of DR, ER and IR motifs with indicated spacing lengths in the binding peaks. **(E)** Distribution of statistically significant motifs of D around the TSS.

The motifs DR[-2] and DR2 were found enriched in both the auxin-mediated promoters and in the DAP-seq and therefore likely represent MpARF1 target motifs. Notably, these motifs are different from the motif preferred by angiosperm ARFs. To investigate potential mechanisms of MpARF1 binding to DR[-2] and DR2 motifs we predicted potential complexes with a combination of RosettaFold2NA (RF2NA; v0.2), a deep learning model trained to predict protein-nucleic acid complexes (Baek *et al.*, 2022), and manual alignments of MpARF1 AlphaFold2 structure (Chapter 3) (Jumper *et al.*, 2021). These predictions suggested that two MpARF1 subunits could bind simultaneously on DR[-2] and DR2 **(Fig. 5)**, although not via the canonical dimerization interface (Chapter 2) (Boer *et al.*, 2014; Kato *et al.*, 2020). Instead, the subunits bound at an angle of 225° (DR[-2]) and 90° (DR2) relative to each other. Notably, for DR[-2] we did not observe interactions between the two subunits **(Fig. 5A, B)**. Given that two TGTCNN sites partially overlap in DR[-2], we indeed observed steric clashes between the MpARF1-H146 loop of one and the R194-R199 loop of the other subunit's B3 domains **(Fig. 5C)**. In contrast, for DR2 we observed that the B3 domains of each subunit bound independently **(Fig. 5E, F)**. This predicted structure created a potential novel interface between Q167 and Q168 of the B3 domain of one subunit, and R320 and E351 of the ancillary domain (AD) of the other subunit **(Fig. 5F)**. Based on previous data (see Chapter 2), we found that the potential interacting residues within the B3 are weakly conserved (ConSurf score 4 and 5 for Q167 and Q168 respectively) although flanked by highly conserved residues **(Fig. 6A)**. In comparison, the residues in the AD domain are more conserved (ConSurf score 7 and 6 for R320 and E351 respectively), although the structural residues following R320 and R351 are highly variable across ARFs **(Fig. 6B)**. These results indicate the predicted interface is not conserved among plants. The Arabidopsis and *M. polymorpha* ARFs mirror these results from Consurf, as Q167 and Q168 are only conserved in AtARF1, AtARF6 and MpARF1 **(Fig. 6C)**. The other A-class and B-class ARFs have different residues in the homologous location of at least one of the two glutamines. In contrast, R320 and E351 are more conserved in the A-class and B-class ARFs. Notably, the C-ARFs only have a conserved E351 site, while the other residues have been replaced. Together, we therefore reveal that an MpARF1 homodimer could bind to a DR2 motif, requiring a novel interaction interface that shows high variability across the plant kingdom.

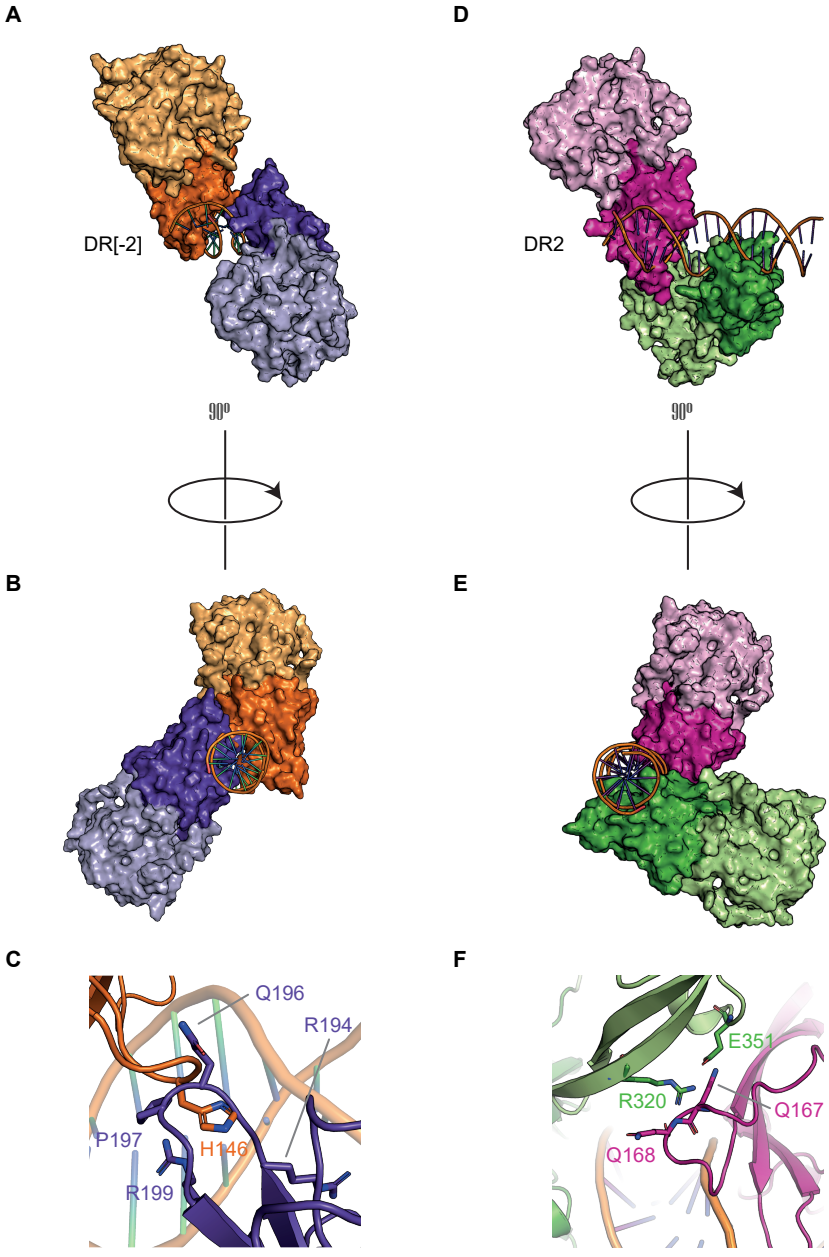


Fig. 5 | Potential binding mechanisms of two MpARF1 subunits on DR[-2] and DR2 motifs. **(A-C)** DR[-2] motif. **(D-F)** DR2 motif. **(A, B, D, E)** DNA shown in cartoon representation, MpARF1 subunits in surface representation. Each subunit in a different colour, with the B3 domains of each subunit in darker shades. **(C, F)** Cartoon represented closeup of: (C) steric clashes between the DNA-binding loops of each subunit in the DR[-2] motif, and (F) potential interface between the B3 domain (Q167 and Q168) and the ancillary domain (R320 and Q351) of each subunit in the DR2 motif.

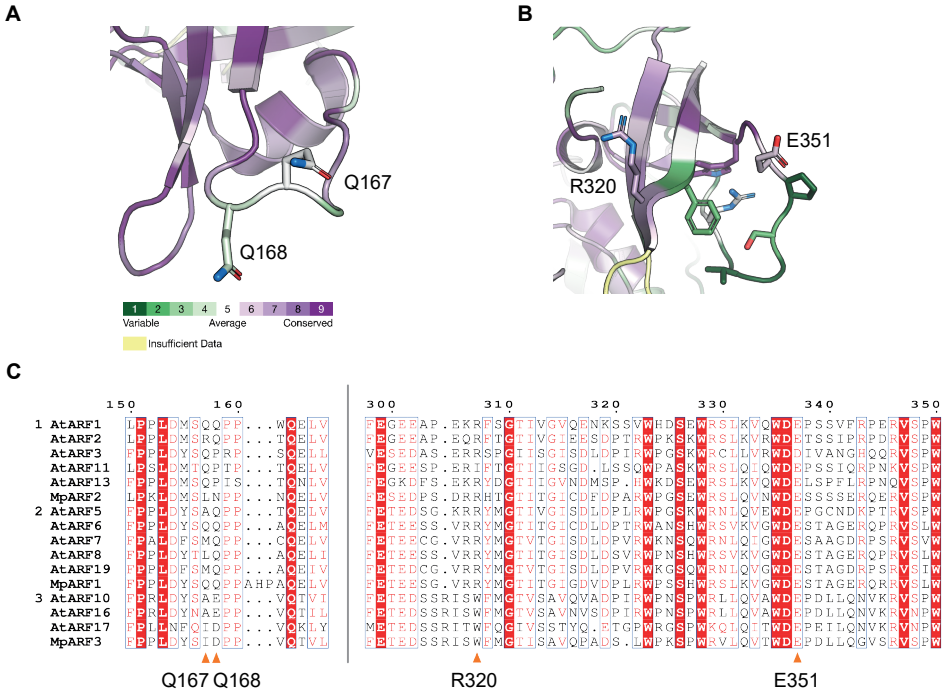


Fig. 6 | Consurf Conservation of residues implicated in the potential homodimerization interface of MpARF1 on DR2 motifs. (A, B) ConSurf analysis (Ashkenazy *et al.*, 2016)(Chapter 3) mapped onto AtARF1 (PDB: 6ycq) with closeup of the potential homodimerization residues of the B3 (A) and AD (B) domains. Interacting residues numbered according to MpARF1 and shown as sticks. ConSurf conservation scores ranges from 1-9 (green to purple), higher is more conserved. Yellow is insufficient data. (B) Residues of a potential hydrophobic pocket behind the interaction residues are also shown as sticks. **(C)** Alignment of Arabidopsis (At) and *M. polymorpha* (Mp) ARFs. Interface residues from A, B and Fig. 5 highlighted with orange triangles.

Discussion

Previous work with *M. polymorpha,* and corroborated by work in *P. patens*, established a model for MpARF1 and MpARF2 competition on shared DNA-binding sites wherein only MpARF1 directly transduces the auxin signal into a transcriptional response. Genetic studies have shown that silencing of Mp*IAA* via RNA interference or increased expression of either Mp*TIR1* or Mp*ARF1* leads to auxin hypersensitivity, whereas non-degradable MpIAA or the null mutants Mp*tir1* and Mp*arf1* are auxin resistant (Flores-Sandoval *et al.*, 2015; Kato *et al.*, 2015; Kato *et al.*, 2017; Sugano *et al.*, 2014; Suzuki *et al.*, 2023). Other factors such as Mp*ARF2*, Mp*ARF3* and Mp*PRAF* – a RAF-like kinase involved in rapid (under two min) auxin responses – have been excluded from direct involvement in the auxin-mediated transcriptional responses on the basis of genetic and molecular studies (Flores-Sandoval *et al.*, 2018a; Flores-Sandoval *et al.*, 2018b; Kato *et al.*, 2020; Kuhn *et al.*, 2022; Mutte *et al.*, 2018). Here, we found that the transcriptome of Mp*arf1-4* is unresponsive to exogenous auxin, thereby supporting that MpIAA and MpARF1 form the basis of the NAP.

When analyzing different auxin-induced transcriptomes we found a minimal overlap between datasets **(Fig. 3B)**, despite some of these having been generated in the same lab. Firstly, this suggests developmental, environmental and even experimental factors may influence the auxin response in such a way that each transcriptome represents a uniquely incomplete fraction of the auxin responses in *M. polymorpha*. Secondly, the use of bulk gemmae could dilute cell type-specific responses wherein auxin plays a key role (e.g., rhizoid initials) since Mp*TIR1*, Mp*ARF1* and Mp*ARF2* show tissue-dependent expression levels (Das *et al.*, 2022). This in turn raises the possibility that there are auxin-responsive tissues in Mp*arf1-4* that are strongly diluted or even missed in bulk gemmae analyses. This could be resolved by performing single-cell RNA-seq (scRNA-seq), which has been used for *M. polymorpha* (Wang *et al.*, 2023). Both scRNA-seq and more bulk auxin-treated transcriptome analyses could improve the resolution of the enrichment for auxin-responsive elements **(Fig. 3)** and could reveal auxin-responsive elements that are associated with particular growth stages or cell types.

Elucidating the DNA-binding preferences of MpARF1 is the key to deepen our understanding of both the least complex NAP and of the evolution of DNA-binding specificities of the ARF family across the plant kingdom. Here, we captured the DNA-binding syntax of MpARF1 using a combination of promoter motif-enrichment analysis of auxin-responsive genes in *M. polymorpha*, DAP-seq and *in silico* structural modelling. We observed that the preferred binding motif for MpARF1 is TGTCGG **(Fig. 4C)**, which was previously shown to be bound *in vitro* (Kato *et al.*, 2020), in line with the binding preferences of angiosperm ARFs (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014; Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). We can therefore conclude that the ARF-B3 binding preferences were already established in the last common ancestor of

land plants.

Our analyses further suggest that MpARF1 preferentially binds DR[-2] and DR2 motifs in whole-genome contexts **(Fig. 3, 4)** and structural modeling suggests that both motif conformations could in principle be bound by two MpARF1 subunits simultaneously **(Fig. 5)**. Binding by two subunits would allow for cooperative binding (i.e. two dimerized transcription factor subunits binding stronger to their target DNA than a single subunit or the sum of two subunits), which is an important aspect of angiosperm ARF binding to IR7 motifs, and indirect evidence suggest it may also be required for binding to DR5 (Boer *et al.*, 2014; Fontana *et al.*, 2023; Freire-Rios *et al.*, 2020; Ulmasov *et al.*, 1997b). For the DR[-2] however, simultaneous binding would require conformational changes in the DNA-binding loops to resolve steric clashes. Since we did not observe any interactions between the two subunits, the DR[-2] motif is likely not bound by cooperative action of the DBDs. This is in contrast to the DR2 motif, where we did observe an interaction between a novel interaction interface. One caveat of the modelling is that we used DBD structures derived from IR7-bound ARFs, which are locked into a specific conformation. Since the B3 domain is flexible (Boer *et al.*, 2014), it is possible that MpARF1 adopts alternative conformations on DR[-2] and DR2. Future biochemical studies need to be focused on whether DR[-2] and DR2 are bound by (cooperative) homodimers. Furthermore, in Arabidopsis, the canonical dimerization helix was established based on crystallography and observations that AtARF5 dimerization mutants failed to bind to IR7 motifs and rescue an *Atarf5* mutant (Boer *et al.*, 2014). Since MpARF1 binding to DR[-2] and DR2 does not seem to rely on the canonical dimerization helix, we predict that canonical dimerization mutants in MpARF1 should be able to (partly) rescue Mp*arf1* null mutants and would still bind DR[-2] and DR2 *in vitro*. On the other hand, Mp*ARF1* mutants for Q167, Q168, R320 or E351 that disrupt putative binding should not be able to rescue the mutant, nor bind *in vitro*. Finally, since DBD-swap experiments have established that MpARF2-DBD functionally only partially overlaps with MpARF1-DBD, testing the binding preferences of MpARF2 would be the next step in elucidating the NAP from *M. polymorpha*.

With the caveat that RNAseq meta-analysis was based on few auxin-responsive genes and that DAPseq was based on a single replicate, it is remarkable that the motifs we found for MpARF1 and in auxin-responsive promoters in *M. polymorpha* are different from those found previously in angiosperms. This begs the question of which motifs were preferred by the ancestral A-ARF. We and others have shown MpARF1 can bind angiosperm-derived motifs DR5 and IR7 **(Fig. 1)** (Kato *et al.*, 2020). However, neither motif is enriched in DAP-seq **(Fig. 4E)**, only IR7 is weakly enriched in a single transcriptome dataset **(Fig. S1)**, and the DR5 reporter is not auxin-responsive in *M. polymorpha* (Ishizaki *et al.*, 2012). In contrast, neither DR[-2] nor DR2 is enriched in promoters of Arabidopsis auxin-induced genes, although DR2 was weakly enriched in

auxin-repressed genes (Freire-Rios *et al.*, 2020). This suggests B-ARFs are binding DR2, but B-ARFs binding peaks were not enriched for DR2 motifs (Galli *et al.*, 2018; Stigliani *et al.*, 2019). Intriguingly however, DR[-2] and DR2 were not enriched in DAP-seq peaks of AtARF5 or its maize homologs ZmARF4 and ZmARF29, but the remaining maize A-ARFs (homologous to AtARF6, 7, 8 and 19) show enrichment for DR2 (Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). Due to differences in nomenclature, the DR[-2] and DR2 in our analysis are DR0 and DR4, respectively, in Galli *et al.* (2018). Given that the predicted DR2 homodimerization interface is variable across ARFs **(Fig. 6)**, ARFs may have different affinities to dimerize and some may be unable to bind DR2 at all. The AtARF5 subfamily originates from a gene duplication event in an ancestral pan-ARF5/7/19 gene in the ancestor of seed plants (Mutte *et al.*, 2018). These results lead to a hypothesis that the last common A-ARF ancestor of terrestrial plants could bind DR2, DR5 and IR7 motifs. In the liverworts, binding to DR2 was favored and binding to DR[-2] gained, while in seed plants, binding to DR5 and IR7 was favored. The AtARF5 subfamily subsequently lost binding to DR2. These hypotheses could be supported or rejected by a cross-species DAP-seq (e.g, MpARF1-DBD challenged with Arabidopsis genomic DNA, and AtARFs with *M. polymorpha* DNA) that could indicate whether intrinsic protein properties or the genomic composition – which was shown to drive B-ARF DNA-binding preferences (Galli *et al.*, 2018) – are driving the motif selection. In conclusion, our results support a hypothesis where diverging evolution in the A-ARF family caused differences in DNA-binding preferences in the liverworts and angiosperms, although the ancestral A-ARF likely shared features of both liverwort and angiosperm ARFs.

Material and methods

Bacterial 1 hybrids

Bacterial 1 hybrids were performed as described (Noyes *et al.*, 2008b). AtARF5-DBD (M1-P390) and MpARF1-DBD (M1-P481) were subcloned from previous clones (Boer *et al.*, 2014; Kato *et al.*, 2017) between the *KpnI* and *XbaI* sites into bait vectors pB1H2ω2 (Addgene #18038), pB1H2ω5 (Addgene #18039) and pB1H2ωL (Addgene #18040), which contain IPTG-inducible promoters of weak, medium and strong induction respectively. A *3xDR5* cassette (TGTCGGcctttTGTCGGcctttTGTCGG) and *IR8/ER13* cassette (TGTCGGtcatgtgaCCGACAaaacacatgatttTGTCGGtcatgtgaCCGACA) were cloned into prey vector pH3U3 (Addgene #18046) between the *EcoRI* and *NotI* sites. Integrity of the inserts, and coding sequences of *ΩRNAP* (pB1H2ω), *HIS3* and *URA3* (pH3U3) were confirmed with Sanger sequencing. AtARF5-R215A and MpARF1-R194A mutants were created via HiFi ssOligo site-directed mutagenesis per manufacturer's instructions (E2621, New England Biolabs). For positive and negative controls, we included *zif268* prey and bait vectors: pH3U3-zif268 (Addgene #18046) and pB1H2ω2-zif268 (Addgene #18045).

E.coli strain DH5α was used for cloning and subcloning. *E.coli* strain USO (Addgene #18049; *hisB- purF- rpoZ-*) was used for B1H screens. USO cells were made Mix&Go competent (T3001, Zymo Research) or freshly made electro-competent as follows. In the late afternoon, inoculate 3 ml of LB with a fresh USO colony and grow at 37°C. The next morning, inoculate 50 ml of sterile LB in a 250 ml sterile Erlenmeyer flask and grow at 37°C (extra competence can be achieved by growing at 22°C). Note that volume of LB can be scaled down or up: calculate 1.5 ml of LB per transformation and maintain a 1:5 or higher ratio of LB to air for proper aeration. Once cells reach an optical density at 600 nm (OD600) of 0.4 to 0.6, transfer 1.5 ml per transformation into a 2 ml Eppendorf tube cooled on ice. Centrifuge 16,000 rcf for 30 sec at 4°C and remove supernatant. Wash cells twice by resuspending with 1 ml of ice-cold sterile deionised water, centrifugation as above and removal of the supernatant. Resuspend the cells in 40 µl MQ and transform the same day. Competent USO cells were generally co-transformed with both prey and bait vectors and selected on LB with Kanamycin and Ampicillin.

To perform B1H assays, overnight cultures of single USO colonies in LB with antibiotics were prepared. The next morning, cultures were normalised to the same OD600 for each culture. Cultures were spotted (5 µl) on NM selective plates (see Noyes *et al.*, 2008) in 10-fold dilution series. Interactions were screened on media without histidine, and positive controls were grown on NM selective plates with additional histidine. Plates were grown at 22°C until colonies were sufficiently grown. Autoactivation was tested by growing USO cells with only bait or only prey vectors on NM media without histidine. Assays were repeated twice starting from new transformants, with identical results.

Transcriptome analysis

Plant material and growth conditions

M. polymorpha plants (accession Tak-1) (Bowman *et al.*, 2017) used to generate RNA-seq dataset 2 and 3 were grown as follows. Gemmae were grown aseptically on solid ½ Gamborg B5 media (hereafter B5; G0209, Duchefa) with 2% sucrose (S0809, Duchefa) following previously described culturing conditions (Chiyoda *et al.*, 2008; Ishizaki *et al.*, 2008; Kubota *et al.*, 2013). Sucrose was added to enhance gemma cup production (Ishizaki *et al.*, 2016). After 5-6 weeks of growth, gemmae were harvested by holding the plate upside down above a tray filled with sterile demineralized water and hammering the bottom of the plate with a metal spoon. Gemmae that fell in the water were collected in a 70 µm cell strainer. For dormant gemmae (dataset 2), the cell strainer was immediately submerged in a solution of liquid ½ B5 media with 1 µM of IAA (I0901, Duchefa) for 1 hour, after which cells were frozen in liquid nitrogen. For non-dormant gemmae (dataset 3), the cell strainer was submerged in liquid ½ B5 media for 3.5 hours, moved to liquid ½ B5 with 1 µM of IAA for 1 hour and frozen in liquid nitrogen. Samples were kept at -70 °C until processing.

RNA-extraction and sequencing

Frozen gemmea were ground into a powder with a bead shaker. RNA was extracted with the RNeasy Plant Mini Kit (74904, Qiagen) including an on-column DNase treatment (69104, Qiagen). RNA-seq library construction was performed with the TruSeq kit (Illumina) and 150-bp paired-end sequencing with BGISEQ-500 was performed by BGI Tech Solutions (Hong Kong). Obtained raw fastq reads were quality controlled using FastQC. Cleaned reads were mapped to the *M. polymorpha* genome (v6.1, https::// marchantia.info/) using Hisat2 for paired-end reads, followed by counting reads per gene with *FeatureCounts* using default parameters. Finally, genes with low counts (<90 across all samples together) were removed and differentially expressed genes were calculated with DESeq2. Datasets 4 and 5 were generated and analysed before in our lab (Kuhn *et al.*, 2022; Mutte *et al.*, 2018).

Bipartite motif enrichment analysis

TGTCNN repeats with different orientations and spacing locations were identified in gene promoter sequences in the *M. polymorpha* genome (v6.1). The notation DR means TGTCNN repeats in the same DNA strands, the repeats in opposite strand are classified into the everted (ER, head-to-head) and inverted (IR, head-to head). The spacing of monomer units starts from an overlap of two bp (this implies the minimal spacer of negative length, -2 bp), such as DR[-2], which is TGTCTGTCNN, and the maximal spacer is +15 bp. For each gene, a promoter included a 3500 bp long region upstream TSS and a downstream region bounded by a minimum length between the total length

of 5' untranslated region (TSS-AUG) and 1,000 bp. We used five RNA-Seq datasets on auxin treatment, two of 10 day old gemmalings treated with 2,4-D (Mutte *et al.*, 2018) or IAA (Kuhn *et al.*, 2022) and three datasets of gemmae treated with IAA (see above and Chapter 5) to test the significance of association between the presence of bipartite TGTCNN elements with specific spacers/orientations in the promoters of genes and its auxin responsiveness (separately, for up- and down-regulation). To estimate the significance of this association we applied the MCOT approach (Levitsky *et al.*, 2019). It assessed the significance of enrichment of dimeric motifs by Fisher's exact test on a contingency table comparing auxin responsive genes (a) and all genes (b) with at least two TGTCNN elements present in the sequence versus auxin responsive genes (c) and all genes (d) with at least one TGTCNN pair in the specific orientation/spacer. The resulting *p*-values were combined with Fisher's method into *meta-p*-value using *metap* R package. Resulting *meta-p*-values were adjusted with Benjamini-Hochberg multiple testing correction procedure.

DAP-seq

Plasmid construction and protein production

The DBD of MpARF1 (M1-P368) was subcloned from a previous clone (Kato *et al.*, 2017) into a modified pIX-HALO vector (Bartlett *et al.*, 2017; O'Malley *et al.*, 2016) via PCR and HiFi Assembly (E2621, New England Biolabs). The pIX-HALO vector was modified to enable HiFi-mediated cloning (in preparation, Pingtao Ding). The plasmid was miniprepped with transfection grade silica columns (740490, Machery-Nagel). For *in vivo* transcription and translation, a 5 µl ALiCE (AL0103050, LenioBio) reaction was prepared with 5 nM vector per the manufacturers instruction, followed by incubation in an orbital shaker at 500 RPM at 25 °C for 48 hrs. Protein production was verified via Western Blot and the protein was used in DAP-seq immediately.

Genomic DNA extraction

Genomic DNA from *M. polymorpha* was extracted using a modified version by (Kasajima *et al.*, 2013) from 14 day old gemmalings. One gram of frozen, ground material was vortexed with 10 ml of PVPP buffer (50 mM Tris-HCl pH9.5, 10 mM EDTA, 4 M NaCl, 1% CTAB, 0.5% PVPP, 1% β-mercaptoethanol) and incubated at 80°C for 30-45 min. Subsequently, 3 ml chloroform:isoamylalcohol (24:1) and 2 ml TE-saturated phenol was added, followed by vortexing and centrifuging at max RPM at room temperature for 20 min. The upper phase was collected, vortexed with 2 ml of chloroform: isoamylalcohol (24:1) and centrifuged at max RPM at room temperature for 10 min. The upper phase was collected and mixed with: 1 volume of MilliQ-water (Millipore), 0.2 volume 3M NaOAc pH5.2 and 2 – 2.5 volume 96% ethanol. The solution was mixed and DNA left to precipitate at -70 °C for 30 min. The DNA was pelleted by centrifugation at 4°C for 1

hr. After a single washing step with ice-cold 70% ethanol, the DNA was redissolved in 300 µl TE with 10 µg/ml RNase A and left for 15 min at room temperature. To DNA was purified and concentrated on the Quagen Genomic-tip 20/G (10223, Qiagen) by mixing the DNA with 1 volume buffer QBT and gravity-flowed through an equilibrated column. The column was washed with 3x 1 ml of buffer QC, followed by elution with 2x 1 ml of buffer QF. The DNA was precipitated with an isopropanol precipitation and dissolved in 200 µl of a 10 mM Tris-Cl pH8.5 solution. Concentration was measured with the Qubit dsDNA BR Assay kit (Q32853, Invitrogen) on a Qubit 4 fluorometer (Q33238, Invitrogen).

Genomic DNA library preparation

Purified *M. polymorpha* genomic DNA was tagmented and Y-adapter ligated using a home-made Tn5 transposase according to (Picelli *et al.*, 2014). Y-adapters from (Bartlett *et al.*, 2017). For tagmentation, multiple reactions were performed in parallel, each consisting of 30 ng DNA mixed with 2.5 µl Tn5 in 5 µl total volume, followed by incubation at 55 °C for 6 min. Tagmentation was quality checked by RT-qPCR as described (Bartlett *et al.*, 2017) with primers JR416 and JR417. Samples were pooled and concentrated with AMPure XP beads (A63880, Beckman Coulter Life Sciences) and DNA concentrations determined with the Qubit dsDNA HS Assay kit (Q32851, ThermoFisher-Scientific).

DAP-seq

ALiCE reactions were combined with 3 µl (bead volume) Magne HaloTag beads (G7281; Promega) into a final volume of 100 µl PBS + 0.005% NP40 (hereafter PBS40). ARF-Halo proteins were allowed to bind to the beads on an orbital shaker at 4 °C and 800 RPM for 1.5 hrs. The beads were captured on a magnetic rack, and washed with 150 µl PBS40 for 3 times. At each wash step, beads were settled on the bottom of the tube by gravity before magnetic capture. After the washing, 16 ng of *M. polymorpha* DNA library was added with 75 µl PBS40 and incubated with the bead-captured proteins at 4 °C and 800 RPM overnight. Unbound DNA was washed away by four washes with 150 µl PBS40 as above. After removing the supernatant at the final wash, the beads were resuspended in 20 µl milliQ water, to which we added 2.5 µl P5 and 2.5 µl P7 (10 µM) PAGE-purified indexing primers (different combinations per sample of JR416-JR427) and 25 µl NEBNext HF 2x mastermix (M0541S; New England Biolabs) was added. Samples were ran in a PCR program of 72°C for 5 min, 98 °C for 30 sec, and 5 cycles of 98 °C 10 sec, 63 °C 30 sec and 72 °C 30 sec. A 5 µl sample of the PCR was used in an RT-qPCR with the Kapa 2x RT mmix (Kapa Biosystems) for 40 cycles, to determine the number of amplification cycles per sample. For the remaining 45 µl of PCR products, the PCR cycles were continued with the same number of cycles required to reach 1/3 of fluorescent saturation in the qPCR. I.e. if a sample required 5 cycles to reach 1/3 of signal in the qPCR, the NEBNext PCR was continued for 5 cycles. Resulting PCR reactions were purified and concentrated with AMPure XP beads. Negative controls for DAP were created by amplifying the Tn5-

prepared DNA library with indexing primers in a NEBNext reaction. Optimal number of cycles were determined with qPCR as above. DNA libraries were pooled and 150-bp paired-end sequencing with the NovoSeq6000 and was performed by GenomeScan (Leiden).

DAP-seq data analysis

The DAP-Seq profile annotation with respect to the genomics regions the peaks overlap has been done using the method from (Boldyreva *et al.*, 2017). Shortly, the start and the finish of each peak have been localized in either CDS, 5' UTR and 3'UTR, 3.5 kb long upstream and downstream regions, or the intergenic spacers. Then the peak weight was split between the covered genomic regions to calculate the resulting coverage percentage. For de novo motif search, we used the STREME tool with default parameters (Bailey, 2021). The foreground sequence set consisted of DAP-Seq peaks, and the background set was generated based on *M. polymorpha* reference genome sequences. Namely, the sequence length distributions for two sets of sequences were kept the same, allowing for a small variation of 1% in the A/T nucleotide content (Raditsa *et al.*, 2023, under review). The position weight matrix (PWM) for the top-ranked overrepresented TGTCGG motif was used for composite element search with the MCOT tool (Levitsky *et al.*, 2019). The composite elements overrepresented in the DAP-Seq profile (DR[-2], DR2, DR6, and ER1) have been mapped to the genomic coordinates to compute their distribution relative to transcription start site positions (see Raditsa *et al.*, 2023, https://github.com/parthian-sterlet/antinoise, additional supporting programs).

Structure predictions

Structural analyses of MpARF1 binding to DR[-2] and DR2 motifs were predicted using a combination of RosettaFold2NA (RF2NA; version 0.2) (Baek *et al.*, 2022), and manual alignments of MpARF1 AlphaFold2 structure (Chapter 3) (Jumper *et al.*, 2021). For DR2 we predicted a complex of the DNA and two MpARF1-B3 domains (P130-N239) flanking a 30 residue long GSS-linker with RF2NA. We then aligned the AlphaFold2 predictions of the entire MpARF1-DBD with each B3 domain using the *align* command in PyMOL Molecular Graphics System (version 2.5.5, Schrödinger, LLC). For DR[-2] we attempted a prediction with DR[-2] motif and the same B3-linker protein as for DR2. As the prediction failed we left only the DR[-2] DNA in the prediction. We then predicted the aligned MpARF1-DBD with a single TGTCGG element with RF2NA and used the *align* command to align the TGTCGG element with each individual TGTC element in the DR[-2] structure, which resulted in correctly aligning the MpARF1-DBDs.

Conservation per residue scores were obtained from the ConSurf analysis performed in Chapter 3. AtARF1 (PDB: 6ycq) was used as a template for a default ConSurf analysis (Ashkenazy *et al.*, 2016).

Supplementary

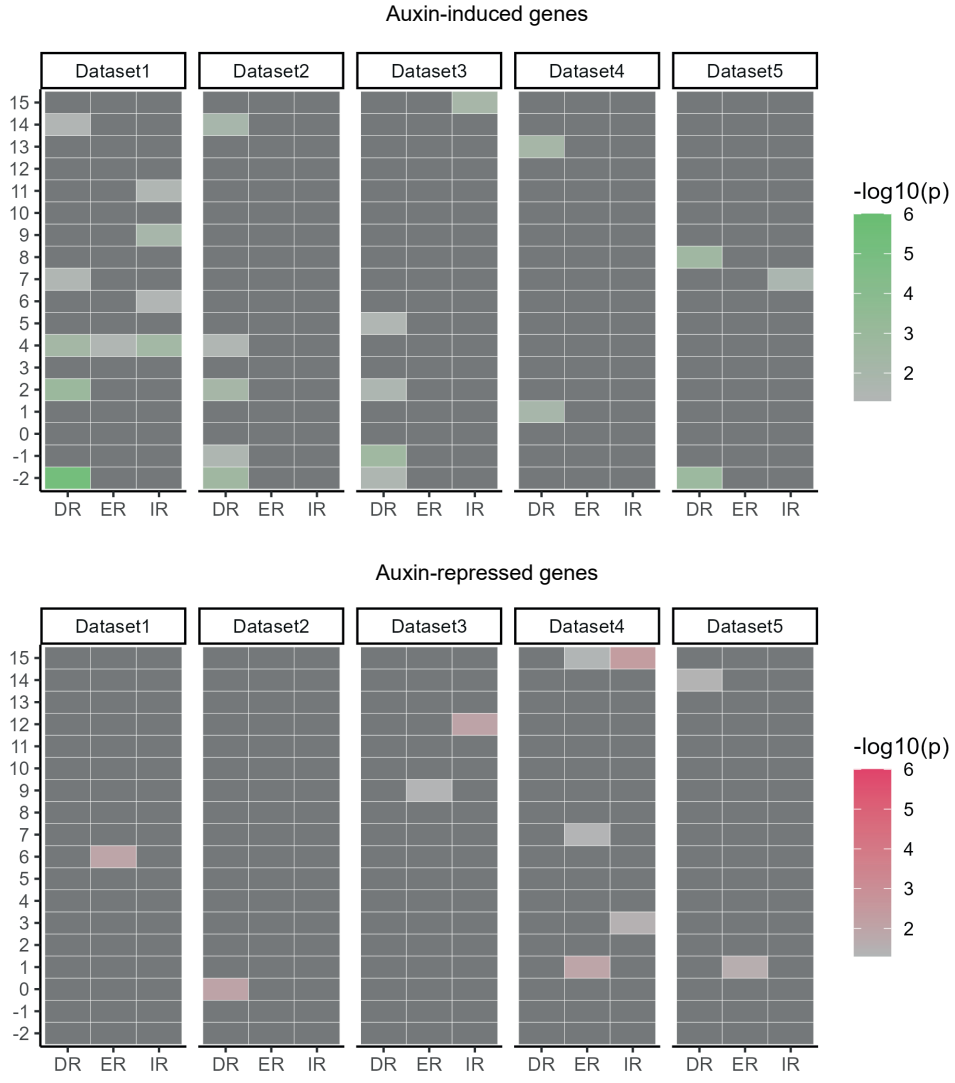


Fig. S1 | Enrichment of motifs in each transcriptome dataset separately. Enrichment of DR, ER and IR motifs with indicated spacing lengths in the upstream (3,500 bp to transcription start site) of the auxin-responsive genes in all datasets. Auxin-induced (green) and auxin-repressed (magenta) genes. Color saturation represents the significance for each bipartite motif in $-\log_{10}$(meta *p*-value). Dataset numbers correspond with Fig. 3A and 3B.

Chapter

5

Evolvability of the ARF-DNA interface

Juriaan Rienstra[1], Sumanth Mutte[1], Martijn de Roij[1], Shubhajit Das[1], Erik van der Velde[1], Yiyun Li[1], Simone van Roosmalen[1], Iraes Rabbers[1], Jorge Hernández-García[1] and Dolf Weijers[1]

[1] Laboratory of Biochemistry, Wageningen University and Research, Wageningen, the Netherlands

Abstract

The AUXIN RESPONSE FACTOR (ARF) transcription factors are found in every lineage of the Streptophytes and exhibit strong conservation of both DNA-binding preferences and DNA-binding residues. ARFs from both the liverwort (*Marchantia polymorpha*) and angiosperm (*Arabidopsis thaliana* and *Zea mays*/maize) show the highest preference for auxin response elements (AuxRE) composed of TGTCGG. The ARF DNA-binding domain (DBD) contains a B3 domain that atomically interacts with these bases, using four residues to contact the core-TGTC (MpARF1-R194, O196, P197 and R199) and one to contact G5 and G6 (MpARF1-H146). Here, we studied the evolvability of the ARF-DNA interface, whether there is any variation for these five DNA-binding residues, and how it affects the MpARF1 function. Variant residues were rarely found for the core-binding residues, and engineered MpARF1 proteins with such variants failed to complement the Mp*arf1-4* mutant. The exception was Q196, for which other variants fully complemented Mp*arf1-4*. In contrast, MpARF1-H146 was highly amendable to other variant residues. Replacing wild-type MpARF1 with variants H146N and H146Y led to delayed growth and low auxin sensitivity caused by lower DNA-binding affinity. Therefore, ARF evolvability is low in the context of DNA-binding residues since most variant residues strongly affect MpARF1's function.

Introduction

The plant-kingdom-specific AUXIN RESPONSE FACTORs (ARFs) are transcription factors (TFs) that are found in every lineage of the streptophytes and form the final step in the auxin signalling pathway (Carrillo-Carrasco *et al.*, 2023; Mutte *et al.*, 2018). In plants, the auxin 3-indole acetic acid (IAA) is synthesized from a tryptophan precursor by successive actions of Tryptophan Aminotransferases (TAA) and YUCCA genes (YUC), followed by signal transduction via the nuclear auxin pathway (NAP). This pathway is composed of three protein families: TIR1/AFB auxin receptors, Aux/IAA transcriptional co-repressors and the ARFs (for a more detailed description see Chapter 2). The TIR1/AFB proteins can bind with auxin, which enables them to bind and degrade Aux/IAA proteins (Tan *et al.*, 2007). The Aux/IAA proteins recruit TOPLESS repressor complexes but cannot bind DNA themselves and instead bind to ARFs, which creates a complex that represses downstream genes (Kato *et al.*, 2020; Tiwari, 2001; Ulmasov *et al.*, 1997b). When TIR1/AFB proteins degrade Aux/IAA's, it allows the ARFs to regulate downstream genes themselves. The last common ancestor of the land plants likely had a single copy of each: a TIR1/AFB, an Aux/IAA and one copy of each of three classes of ARFs. Indeed, in each extant land plant the ARF family can be divided into three classes: A-ARFs, B-ARFs and C-ARFs (Mutte *et al.*, 2018; Okushima *et al.*, 2005). Studies with the liverwort *Marchantia polymorpha*, which mirrors the predicted ancestral state with a single copy of each gene of the NAP, revealed that only the A-ARF (Mp*ARF1*) switches between repressive state when recruiting the Aux/IAA (Mp*IAA*) and an activating state with increasing auxin concentrations (Kato *et al.*, 2015; Kato *et al.*, 2020). The B-ARF (Mp*ARF2*) can recruit TOPLESS complexes by itself and competes with MpARF1 on similar DNA-sites *in vivo*, likely to attenuate auxin signalling indirectly (Kato *et al.*, 2020). The C-ARF (Mp*ARF3*) has been implicated only indirectly, but likely is not involved with auxin signalling directly (Flores-Sandoval *et al.*, 2018a; Flores-Sandoval *et al.*, 2018b; Mutte *et al.*, 2018). Finally, feedback loops occur within the NAP, since MpARF1 was shown to regulate auxin biosynthesis genes Mp*TAA* and Mp*YUC1*, and might regulate Mp*IAA* (Flores-Sandoval *et al.*, 2015; Mutte *et al.*, 2018). In other words, the TIR1/AFB (Mp*TIR1*), Mp*IAA* and Mp*ARF1* form the core auxin signalling components in *M. polymorpha*.

Since ARFs decide which genes are under control of auxin by binding their promoter regions, studying ARF DNA-binding preferences has been key in unravelling the NAP. The DNA-binding domain (DBD) of the ARFs contain a B3 domain that makes atomic contacts with an auxin response element (AuxRE) composed of the nucleotides TGTCNN (Boer *et al.*, 2014). The TGTC form the core sequence, as mutations disrupt DNA-binding (Ulmasov *et al.*, 1997a, 1999b; Ulmasov *et al.*, 1995), and the final two nucleotides (N5 and N6) can be variable, as seen in both ARF DNA-binding assays and *in vivo* (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014; Lieberman-Lazarovich *et al.*, 2019; Zemlyanskaya *et al.*, 2016). Indeed, promoter enrichment analysis in the angiosperm *Arabidopsis thaliana*

(Arabidopsis) showed particular TGTCNN elements to be associated with specific processes, e.g.: TGTCGG with cell-wall related genes and TGTCTT with the phosphate response (Lieberman-Lazarovich *et al.*, 2019; Zemlyanskaya *et al.*, 2016). Nevertheless, DNA-binding assays with A-ARF and B-ARFs from angiosperms *Arabidopsis thaliana* and maize (*Zea mays*) showed that all tested ARFs have the highest affinity for TGTCGG elements (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014; Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019), and we have previously shown that this preference is shared by the liverwort *Marchantia polymorpha* MpARF1 (Chapter 4). Even C-ARFs bind TGTCGG elements (Martin-Arevalillo *et al.*, 2019), although it is unknown whether this represents the highest affinity element for C-ARFs, thereby indicating strongly conserved DNA-binding preferences within the ARF family. This is supported by crystallography and structural modelling that shows strong conservation of the atomic mechanisms through which the ARF-B3 domains bind to AuxREs (Chapter 3) (Boer *et al.*, 2014; Mutte *et al.*, 2018). The residues that bind to the AuxRE can be broadly divided into residues that: i) confer general DNA affinity by binding the DNA phosphate backbone, and ii) confer AuxRE specificity by entering the major groove. Out of the second group, using Arabidopsis AtARF1 as a reference, four residues bind to the core TGTC: AtARF1-R181, Q183, P184 and R186; and one residue binds the variable N5 and N6 nucleotides: AtARF1-H136 (Chapter 3) (Boer *et al.*, 2014). Comparison between crystal structures of ARFs bound with the medium affinity TGTCTC and high affinity TGTCGG AuxRE have highlighted that conformational flexibility of AtARF1-H136 can affect affinity (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020). These DNA-binding residues are strongly conserved in the land plant ARFs, even between the three ARF classes, as well as in ARFs from extant charophyte algae (Chapter 3) (Martin-Arevalillo *et al.*, 2019; Mutte *et al.*, 2018). However, the homologous residues in MpARF1 have yet to be functionally validated. Nevertheless, ARFs show a strongly conserved DNA-binding interface and DNA-binding preferences for TGTCGG elements.

In this chapter, we explored the evolvability of the ARF DNA-binding interface: whether variation for DNA-binding residues exists within the ARF family, and what their effects could be on the functioning of an ARF. From the relatively simple ancestral state, the ARF family largely expanded in the angiosperms, which could have led to sub- and neofunctionalization of ARFs. This process has been accompanied by changes in DNA-binding preferences in other TF families (Anderson *et al.*, 2015; Badis *et al.*, 2009; Baker *et al.*, 2011; Berger *et al.*, 2008; Sayou *et al.*, 2014), and in ARFs for example, variation for AtARF1-H136 could shift which TGTCNN elements are bound with high or medium affinities. Therefore, we investigated variation for DNA-binding residues within the ARF family and assessed their effect on the function of MpARF1 in *M. polymorpha*. The liverwort has the advantage of a relatively straightforward NAP, i.e. observing the effect of changes in DNA-binding residues of MpARF1 could not be confounded by redundancy with other ARFs. We found variation for TGTC-core binding residues is detrimental for

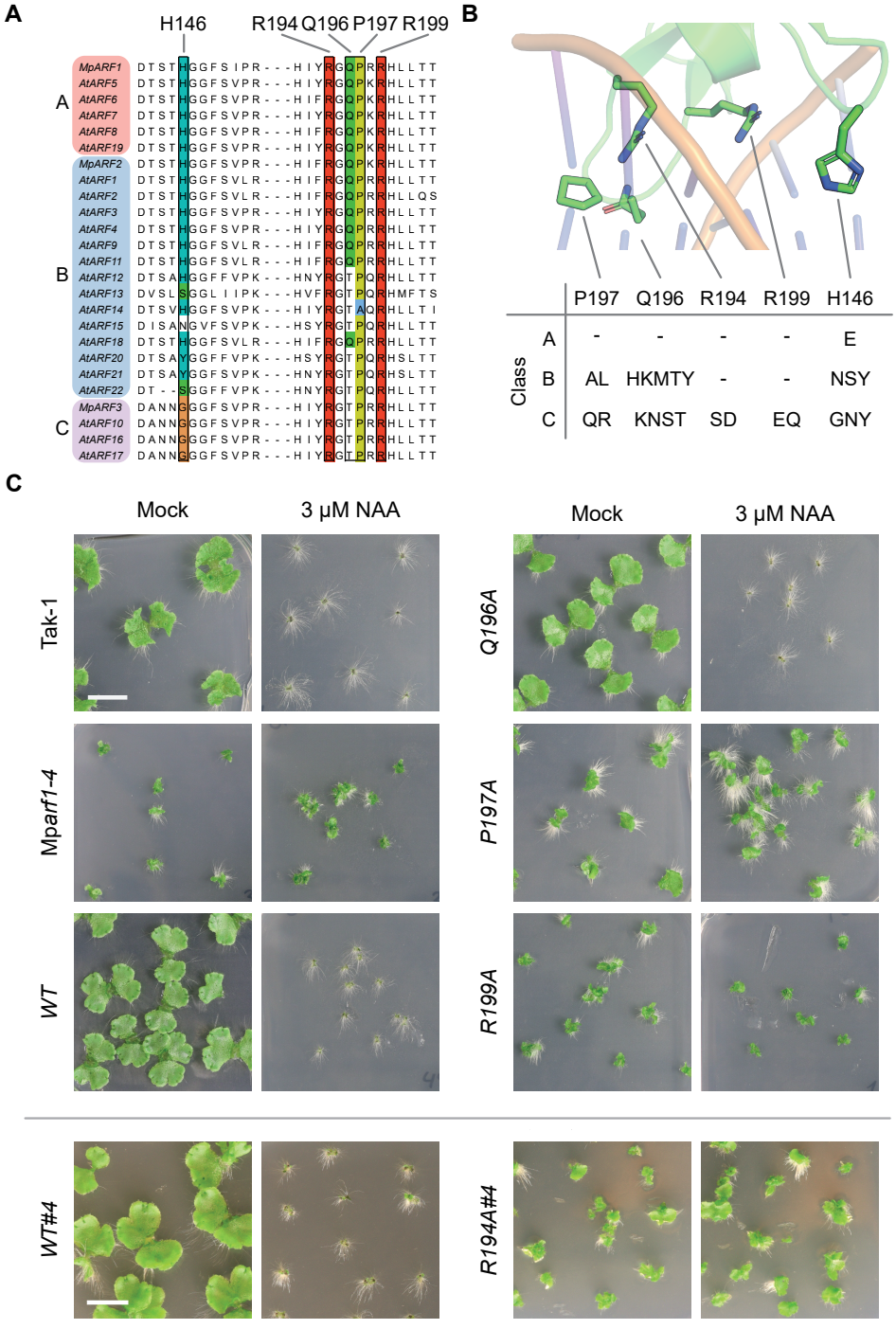
function, while variation on MpARF1-H146 (homologous to AtARF1-H136) was surprisingly tolerable. Using transcriptomics we showed how feedback loops in MpARF1-H146 variants were affected, and used molecular docking to explain the underlying atomic mechanisms for altered DNA-binding.

Results

Variation for ARF DNA-binding residues

To find whether variation for the ARF DNA-binding residues exists, we investigated three levels of evolution: variation between different accessions of *Marchantia polymorpha*, between ARFs from the expanded family of Arabidopsis, and within the plant kingdom. Hereafter, we refer to DNA-binding residues based on MpARF1 unless stated otherwise: R194, Q196, P197 and R199 that bind the core TGTC and H146 that binds the variable N5 and N6 nucleotides. In 128 sequenced *Marchantia* accessions, sampled across the world, we found no variation at all for DNA-binding residues in MpARF1, MpARF2 or MpARF3 (personal communications, Pierre-Marc Delaux). This highlights strong purifying selection in the *Marchantia* genus, and suggests that its adaptation to different environments did not select for alternative DNA-binding of the ARFs. Compared to Marchantia with only one of each class of ARF, Arabidopsis has a more expanded ARF family (Mutte *et al.*, 2018). Nevertheless, within the A-ARF and C-ARF clades, all paralogs have the same DNA-binding residues **(Fig. 1A)**. Note that C-ARFs have a conserved Gly (G) and Thr (T) at position H146 and Q196 respectively. For the B-ARFs, we observed alternative DNA-binding residues in a subclade of B-ARFs (AtARF12, 13, 14, 15, 20, 21 and 22) that are clustered near the centromere of chromosome 1 and are likely recent duplications **(Fig. 1A)** (Okushima *et al.*, 2005). In this subclade, all ARFs have Q196 replaced with Thr, H146 by Asn (N), Ser (S) and Tyr (Y), and AtARF14 has P197 replaced to Ala (A). Neither R194 nor R199 is mutated in this group of B-ARFs. These results highlight that mutations in DNA-binding residues do occur in paralogs, which implies neo -or sub-functionalization of ARFs might be associated with alternative DNA-binding mechanisms.

Fig. 1 | Variation in MpARF1 for major groove binding residues and phenotypic results. **(A)** Sequence alignment of ARF DNA-binding loops from *Arabidopsis thaliana* and *Marchantia polymorpha* (AtARFs and MpARFs respectively). DNA major-groove binding residues are colored according to physiochemical properties (cyan, aromatic; green, polar; red, positive charge; orange, glycine; yellow, proline). **(B)** Tabulation of variation for DNA-binding residues of MpARF1 in the ARF phylogenetic tree from Mutte *et al.* (2018). Crystal structure shown as reference is AtARF1 (PDB: 6ycq). **(C)** Fourteen-day old gemmalings from Takaragaike-1 (Tak-1), Mp*arf1-4* and Mp*arf1-4* transformed with $_{pro}$Mp*ARF1:*Mp*ARF1* constructs without (*WT*) or with indicated mutations grown on medium without (mock) and with 3 µM NAA. Upper panel, shown are representative lines, out of 4 (*WT*) to 8 (mutants) independent transgenic lines. Lower panel, shown are lines *WT#4* and *R194A#4*, representative of all *WT* (4) and *R194A* (3) lines. Note that lower panel are repeated in Fig. 3. Scale bar is 10 mm.


A
H146
R194 Q196 P197 R199
MpARF1 DTSTHGGFSIPR---HIYRGQPRRHLLTT
AtARF5 DTSTHGGFSVPR---HIYRGQPKRHLLTT
AtARF6 DTSTHGGFSVPR---HIFRGQPKRHLLTT
AtARF7 DTSTHGGFSVPR---HIYRGQPKRHLLTT
AtARF8 DTSTHGGFSVPR---HIFRGQPKRHLLTT
AtARF19 DTSTHGGFSVPR---HIYRGQPKRHLLTT
MpARF2 DTSTHGGFSVPR---HIFRGQPKRHLLTT
AtARF1 DTSTHGGFSVLR---HIFRGQPRRHLLTT
AtARF2 DTSTHGGFSVLR---HIFRGQPRRHLLQS
AtARF3 DTSTHGGFSVPR---HIYRGQPRRHLLTT
AtARF4 DTSTHGGFSVPR---HIYRGQPRRHLLTT
AtARF9 DTSTHGGFSVLR---HIFRGQPRRHLLTT
AtARF11 DTSTHGGFSVLR---HIFRGQPRRHLLTT
AtARF12 DTSAHGGFFVPK---HNYRGTPQRHLLTT
AtARF13 DVSLSGGLIIPK---HVFRGTPQRHMFTS
AtARF14 DTSVHGGFSVPK---HIYRGTAQRHLLTI
AtARF15 DISANGVFSVPK---HSYRGTPQRHLLTT
AtARF18 DTSTHGGFSVLR---HIFRGQPRRHLLTT
AtARF20 DTSAYGGFFVPK---HSYRGTPQRHSLTT
AtARF21 DTSAYGGFSVPK---HNYRGTPQRHSLTT
AtARF22 DT--SGGFFVPK---HNYRGTPQRHLLTT
MpARF3 DANNGGGFSVPR---HIYRGTPRRHLLTT
AtARF10 DANNGGGFSVPR---HIYRGTPRRHLLTT
AtARF16 DANNGGGFSVPR---HIYRGTPRRHLLTT
AtARF17 DANNGGGFSVPR---HIYRGTPRRHLLTT
A
B
C
B
P197 Q196 R194 R199 H146
Class
A - - - - E
B AL HKMTY - - NSY
C QR KNST SD EQ GNY
C
Mock
3 μM NAA
Tak-1
Mparf1-4
WT
WT#4
Mock
3 μM NAA
Q196A
P197A
R199A
R194A#4

To understand which variation for DNA-binding residues occurs within the plant kingdom, we tabulated all alternative DNA-binding residues in the phylogenetic tree reported in Mutte *et al.* (2018) which was composed of 258 A-ARFs, 280 B-ARFs and 123 C-ARFs covering all Streptophyte clades (see also: https://itol.embl.de/tree/1372242521 3318111502973365). These results are summarized in **Fig. 1B**. Within the A-ARFs, variation is rare and we only found a single Glu (E) variant on position H146, which was found in one of the three *AtARF6* homologues in the Magnoliid *Calycanthys floridus*. Compared with the A-ARFs, the B-ARFs showed more variation for the DNA-binding residues. For H146, besides the variants found in the Arabidopsis B-ARFs discussed above, only a single B-ARF from the angiosperm *Hakea prostrata* has a Glu (E) variation. Q196 is the most varied position, with 19% (53) of the B-ARFs having a His at this position. All of these B-ARFs belong to the mosses, lycophytes and ferns and are not limited to a single B-ARF clade, indicating independent mutations are causal for the variant. The second common variant is a Thr, found only in the Arabidopsis B-ARFs. Finally, C-ARFs have a conserved Gly at the H146 position **(Fig. 1A)** (Mutte *et al.*, 2018), and no C-ARF had a His at this position. We found variation on all the DNA-binding residues, but these were all concentrated in four C-ARFs: *OsARF20*, a likely pseudogene with two B3 domains (Shen *et al.*, 2010; Xu *et al.*, 2019; Yang *et al.*, 2019), and maize (*Zea mays*) *ZmARF31*, *ZmARF32* and *ZmARF33*. Together, while variation for DNA-binding residues exists in the plant kingdom, it is rare and often found in species with expanded ARF families, highlighting the strong conservation of DNA-binding by the ARFs.

The core TGTC-binding residues are functionally critical for MpARF1

Four out of the six residues that enter the major groove to confer sequence-specificity bind to the core TGTC (see Chapter 3): MpARF1-R194, Q196, P197 and R199. To investigate whether these residues were functionally important for MpARF1, we performed an alanine (A) scanning experiment. As was observed previously (Ishizaki *et al.*, 2012; Kato *et al.*, 2015; Kato *et al.*, 2017; Sugano *et al.*, 2014), the synthetic auxin 1-Naphthaleneacetic acid (NAA) triggers severe inhibition of growth and overprofileration of rhizoids for Tak-1 or Mp*arf1-4* mutants transformed with a wild type $_{pro}$Mp*ARF1:*Mp*ARF1* construct (hereafter Mp*ARF1-WT* or *WT*), whereas the Mp*arf1-4* mutant does not respond and remains small in both mock and auxin conditions **(Fig. 1C)**. Gemmalings from Mp*arf1-4* transformed with $_{pro}$Mp*ARF1*:Mp*ARF1-R194A*, *P197A* and *R199A* (hereafter Mp*ARF1-R194A*, Mp*ARF1-P197A*, etc.) exhibited Mp*arf1-4*-like inhibition of growth and lack of rhizoids on NAA. This indicated a failure to complement and thereby the functional importance of MpARF1-R194, P197 and R199 **(Fig. 1C)**. In contrast, Mp*arf1-4* transformed with Mp*ARF1-Q196A* was fully complemented, which makes its exact role in DNA-binding unclear. In the crystal structure (PDB: 6ycq), the homologous AtARF1-Q183 bonds via its backbone with nucleobase C2' while the side chain is rotated outwards towards the DNA phosphate backbone without forming hydrogen bonds in the crystal structure **(Fig.**

2A, Chapter 3) (Boer *et al.*, 2014). However, we did observe that different Q183 rotamers could potentially form a hydrogen bond with the phosphate groups of nucleotide A3' **(Fig. 2A)**. Since the backbone, rather than the side chain, confers sequence specificity, it is likely any other residue could functionally replace Q196. To test this, we transformed Mp*arf1-4* with *Q196K* and *Q196R* constructs, both which fully complemented **(Fig. 2B)**. This confirms other residues can functionally replace Q196, thereby explaining the high variability for this position in the B-ARFs and C-ARFs **(Fig. 1B)**.

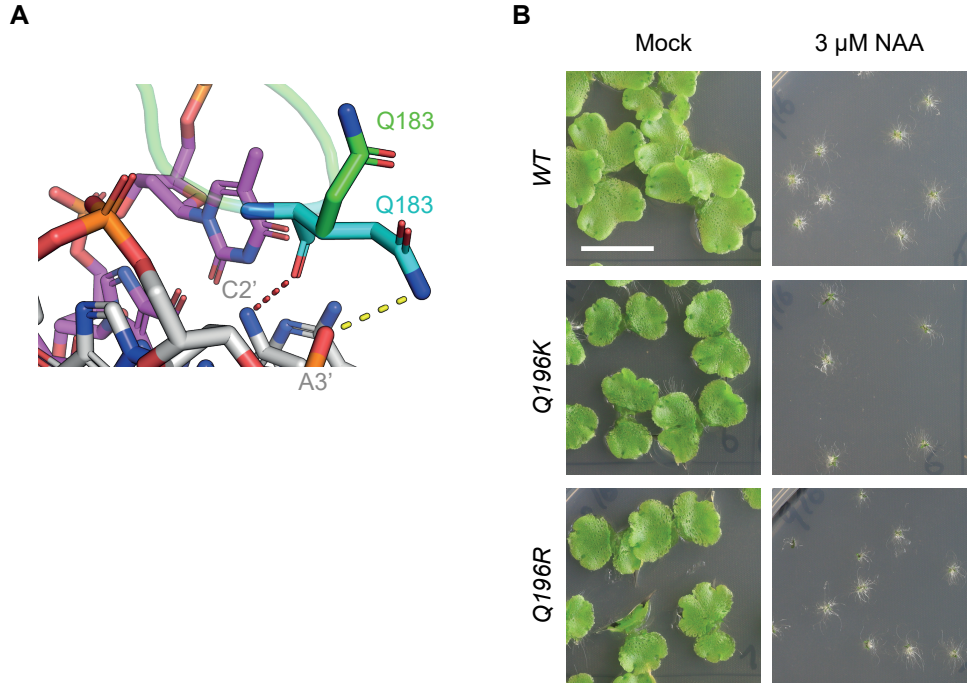


Fig. 2 | MpARF1-Q196 closeup and phenotypic results. (A) Closeup of interaction between AtARF1-Q183 (cyan and green; homologous to MpARF1-Q196) and DNA (white, magenta). Q183 as oriented in the crystal structure (green) and in alternative rotamer (cyan). Dashed lines represent hydrogen bonds, with red hydrogen bond formed irrespective of Q183 rotamer, and yellow only in the rotamer (cyan). PDB: 6ycq. **(B)** Fourteen-day old gemmalings of Mp*arf1-4* complemented with Mp*ARF1-WT*, *Q196K* and *Q196R* construct. Shown are representative lines, out of 4 (*WT*) to 8 independent transgenic lines.

For the *R194* and *R199* positions, we found no variation in the A-ARFs and B-ARFs. Among the proteinogenic amino acids, Arg (R) is unique for its length and for the capacity to form multiple simultaneous hydrogen bonds, as illustrated by the ARFs, where R194 and R199 form extensive hydrogen bond networks between the B3 domain, water molecules and the DNA (Chapter 3) (Boer *et al.*, 2014). The only amino acid with similar physiochemical properties is Lys (K), which is shorter and has less capacity to hydrogen bond. Additionally, we found C-ARFs with negatively charged amino acids – Glu (E) and

Asp (D). We therefore complemented Mp*arf1-4* with Lys (R194K, R199K) and Glu (R199E) substitutions **(Fig. 3)**. Compared to Mp*arf1-4* or the *R194A* gemmalings, the *R194K* gemmalings were slightly bigger, indicating the construct can partially complement and therefore partially retains DNA-binding capacity for TGTCNN elements. For *R199* in contrast, neither *R199A*, *R199E* nor *R199K* constructs were able to complement Mp*arf1-4* **(Fig. 3)**. These results show that Arg fulfills a crucial role in ARF DNA-binding that other amino acids cannot fulfil.

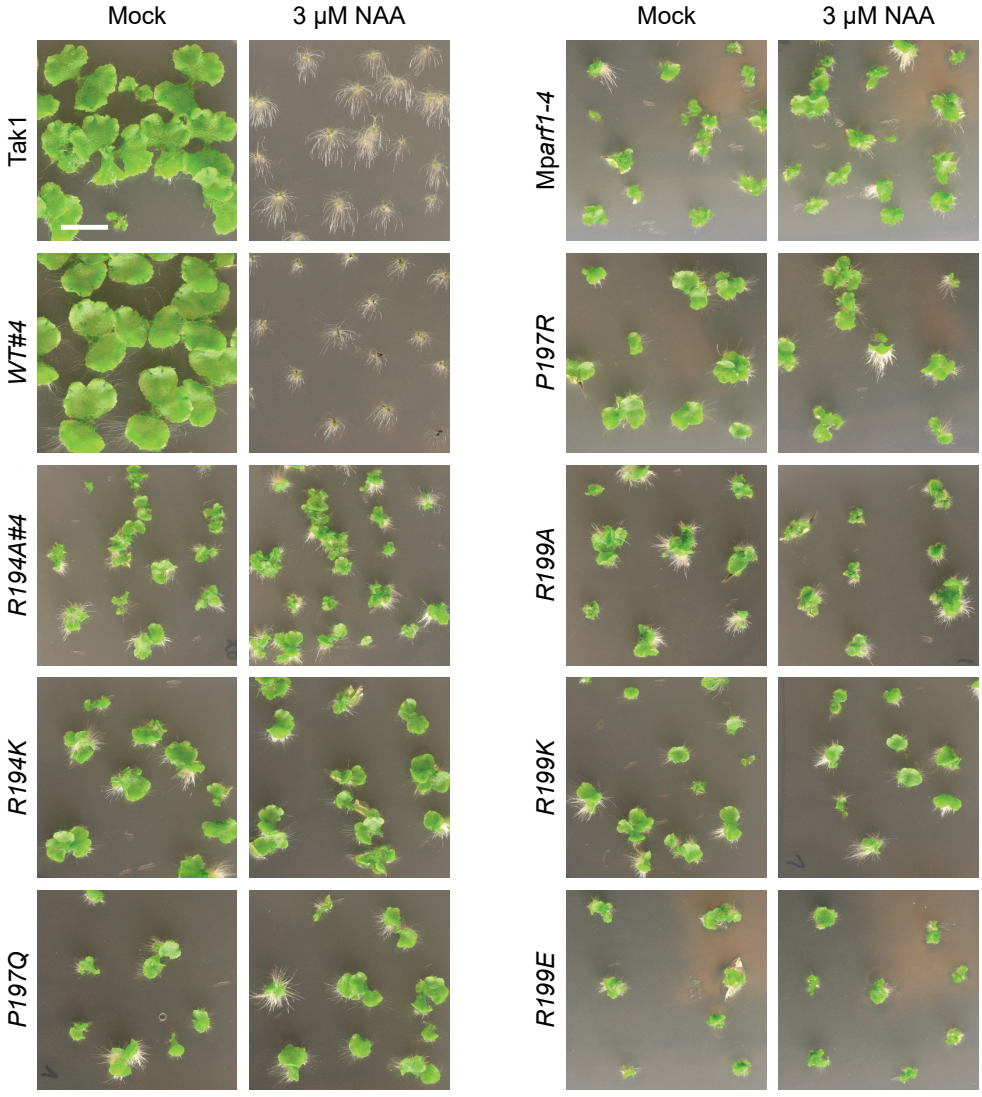


Fig. 3 | Growth phenotypes of AuxRE core-binding variants. Fourteen-day old gemmalings from Tak-1, Mp*arf1-4* and Mp*arf1-4* transformed with Mp*ARF1-WT* (*WT#4*) or with indicated mutations grown on medium without (mock) and with 3 µM NAA. Scale bar is 10 mm.

P197 plays a dual role in ARF DNA-binding: forming hydrophobic contacts with the methyl-group of DNA base T1 (Chapter 3) (Boer *et al.*, 2014), and performing a structural role due to proline's sterical limitations (discussed in Chapter 3). Since the *P197A* constructs failed to complement **(Fig. 1C)** – where Ala (A) is a short, hydrophobic residue with more *phi* and *psi* flexibility compared to Pro – it showed its structural role is likely crucial. Nevertheless, we observed variants with long, polar residues **(Fig. 1B)** and therefore tested the mutants *P197Q* and *P197R*. Neither construct complemented Mp*arf1-4* however **(Fig. 3)**, thereby indicating that *P197* is not easily functionally replaced.

To test whether the failure to complement by *R194*, *P197* and *R199* constructs was due to protein expression or stability problems, we analyzed correct localization of the proteins in the gemmae. Since all core-binding variants discussed above (except *R194A*, see material and methods for an overview) were $_{pro}$Mp*ARF1:*Mp*ARF1-Citrine* fusion proteins, a fusion that does not affect complementation by Mp*ARF1* **(Fig. S1)**, we could determine their subcellular localization. In *WT* lines, MpARF1-Citrine is broadly present in gemmae and is primarily nuclear localized **(Fig. 4)**. This matches expression patterns of endogenous Mp*ARF1* fused to fluorescent tags via homologous recombination (Das, 2022), indicating that our construct can recapitulate the endogenous expression of Mp*ARF1*. This pattern was also observed in both complementing (Q196A) and non-complementing lines **(Fig. 4)**, thereby indicating the mutations do not cause misexpression. Additionally, all mutant proteins were nuclear localized and showed similar protein levels as the *WT* lines based on Citrine intensities **(Fig. 4)**, thereby

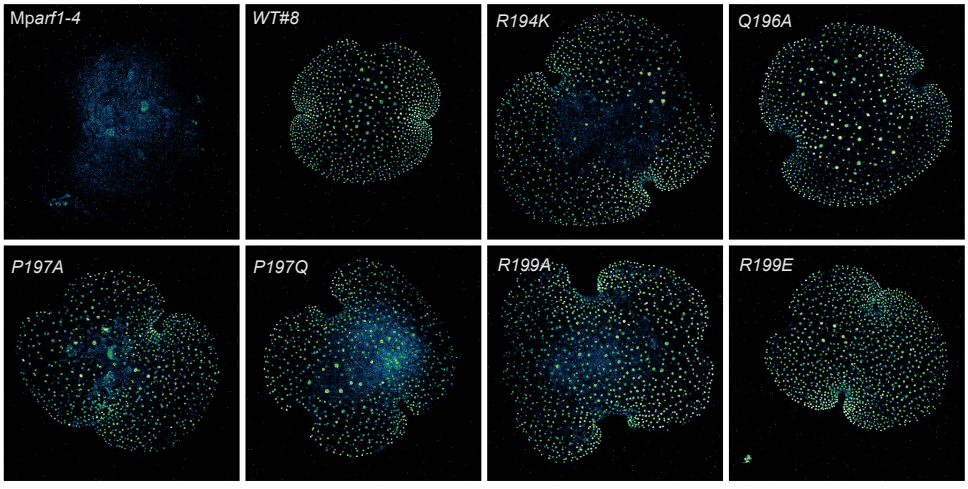


Fig. 4 | MpARF1-Citrine fusion protein localization and expression in gemmae. Gemmae of Mp*arf1-4* transformed with $_{pro}$Mp*ARF1:*Mp*ARF1-Citrine* with indicated mutations were imaged. Genotype is indicated in the corner of each image, where each image is representative of at least 3 gemmae. Fluorescence signal is scaled from low to high, blue to yellow.

indicating the constructs result in similar protein levels and nuclear localization as *WT* constructs. We thereby confirm the previous phenotypes of the mutant lines were caused by issues with DNA-binding.

AuxRE core-binding residues have different DNA-binding mechanisms

Many variants for the residues binding to the TGTC-core of an AuxRE were unable to rescue the Mp*arf1-4* mutant **(Fig. 1, 2, 3)**. To understand if these variants were affected in DNA-binding, we performed *in silico* molecular docking using the crystal structure of Mp*ARF2* (PDB: 6sdg) as template **(Fig. 5)**. The MpARF1-R194K variant was weakly

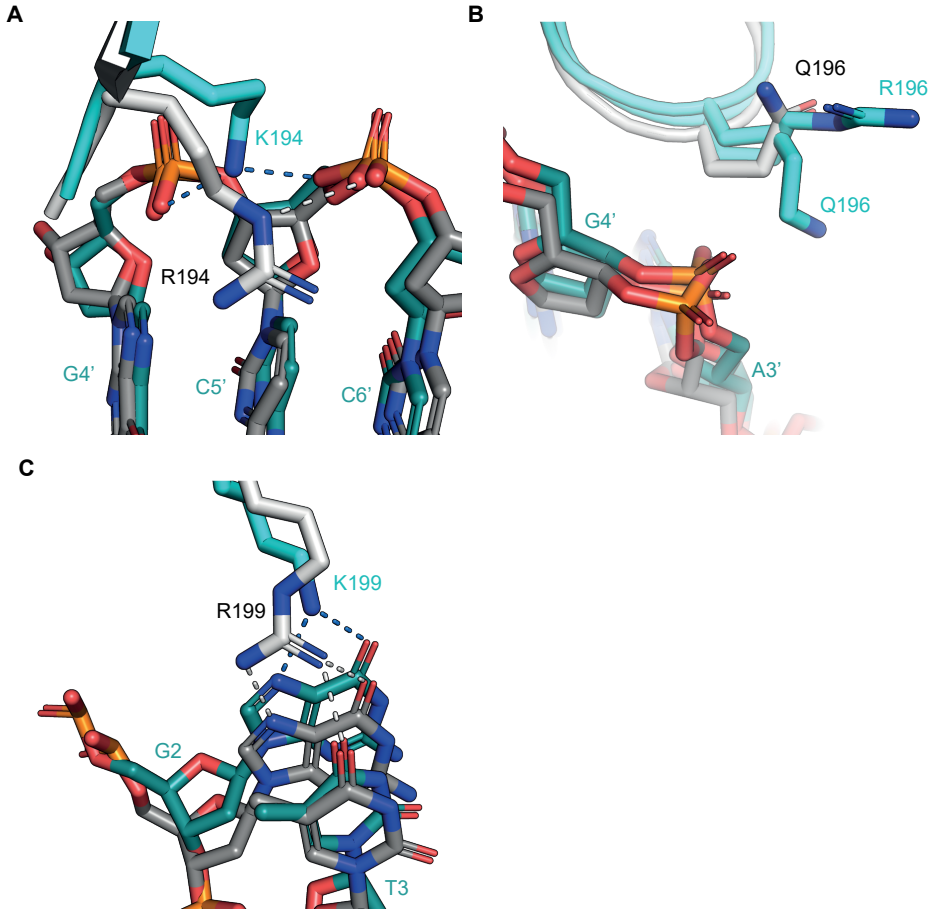


Fig. 5 | Molecular docking of MpARF1-R194, Q196 and R199. Shown are interfaces of R194 **(A)**, Q196 **(B)** and R199 **(C)** before and after molecular docking. Note MpARF2 with TGTCGG DNA (PDB: 6sdg) was used for molecular docking, but is referred as MpARF1. In all, white and grey colours indicate wild type ARF and DNA respectively. Cyan and teal indicate docked structures of specific variants indicated in each panel. Dashed lines represent hydrogen bonds in the wild type (white) and docked (marine) structures.

complementing **(Fig. 3)**, and while R194 can enter the major groove to confer base specificity, K194 only reaches and hydrogen bonds with the phosphate backbone **(Fig. 5A)**. Given that each pair of nucleotides has different physiochemical properties (e.g., width, tilt, roll; Adam *et al.*, 2019), it is possible K194 is positioned in such a way to confer base-specificity simply by the distance and angles of the G4' and G5' phosphate groups, although not sufficient to fully complement. The Q196 variants were all able to complement Mp*arf1-4* **(Fig. 1, 2)**, which can be explained by the docking. None of the amino acid side-chains from Q196, K196 or R196 makes contacts with the nucleotides, instead rotating outwards towards the solution **(Fig. 5B)**. Finally, neither R199A nor R199K was able to complement Mp*arf1-4* **(Fig. 3)**. In the docking, both R199 and K199 were able to hydrogen bond to nucleobase G2 **(Fig. 5C)**, with R199 additionally hydrogen bonding with T3 **(Fig. 5C)**. K199 therefore loses part of the base specificity. Although *in silico* docking does not show water molecules, it is likely both K194 and K199 do not have the same capacity as R194 and R199 to correctly place water molecules between the B3 domain and DNA. Altogether, the docking explains why R194 and R199 are irreplaceable while Q196 is more amendable.

MpARF1-H146 mutants are partially complementing

The AtARF1-H136 residue determines whether a DNA element is a high affinity (TGTCGG) or medium affinity (TGTCTC) site (Chapter 2) (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020), and the observed variation for this residue therefore could have two effects **(Fig. 1A, 1B)**. Either the variant could shift the distribution of which TGTCNN sites are high and medium affinity, and/or it could reduce or increase DNA-binding affinity to TGTCNN sites overall. Either way, given that in Arabidopsis, different TGTCNN sites are associated with specific processes and cell-types (Lieberman-Lazarovich *et al.*, 2019; Zemlyanskaya *et al.*, 2016), we speculated that H146 variants could show altered developmental phenotypes in *M. polymorpha.* We therefore complemented Mp*arf1-4* with *MpARF1* constructs with the following mutations: *H146A*, *H146G, H146N*, *H146S*, *H146Q*, and *H146Y.* The non-complementing variant *R194A* **(Fig. 1C, 3)** was used as a negative control. Gemmalings of recovered transgenic lines were grown for 14 days and their projected thallus area measured **(Fig. 6A)**. As expected, Mp*arf1-4* and the *R194A* lines showed impaired growth. Notably, all the H146 variants complement the impaired growth of Mp*arf1-4* to a certain extent, with most variants having comparable thallus area to Tak-1 and *WT* lines – i.e., *H146A*, *H146G*, *H146Q*, and *H146S* – indicating full complementation. One of the WT lines, *WT#2*, had a thallus area 0.5 times of the other two *WT* lines and Tak-1, a similar area as all the *H146Y* lines and two of the *H146N* lines. While *H146N#1* and *H146N#2* have a reduced area, *H146N#3* is fully complementing.

Since a previous study showed that Mp*ARF1* overexpression lines produced more gemma cups (Eklund *et al.*, 2015; Flores-Sandoval *et al.*, 2015), we tested if the *H146* mutants were affected in gemma cup production. Each line was grown for 5 weeks

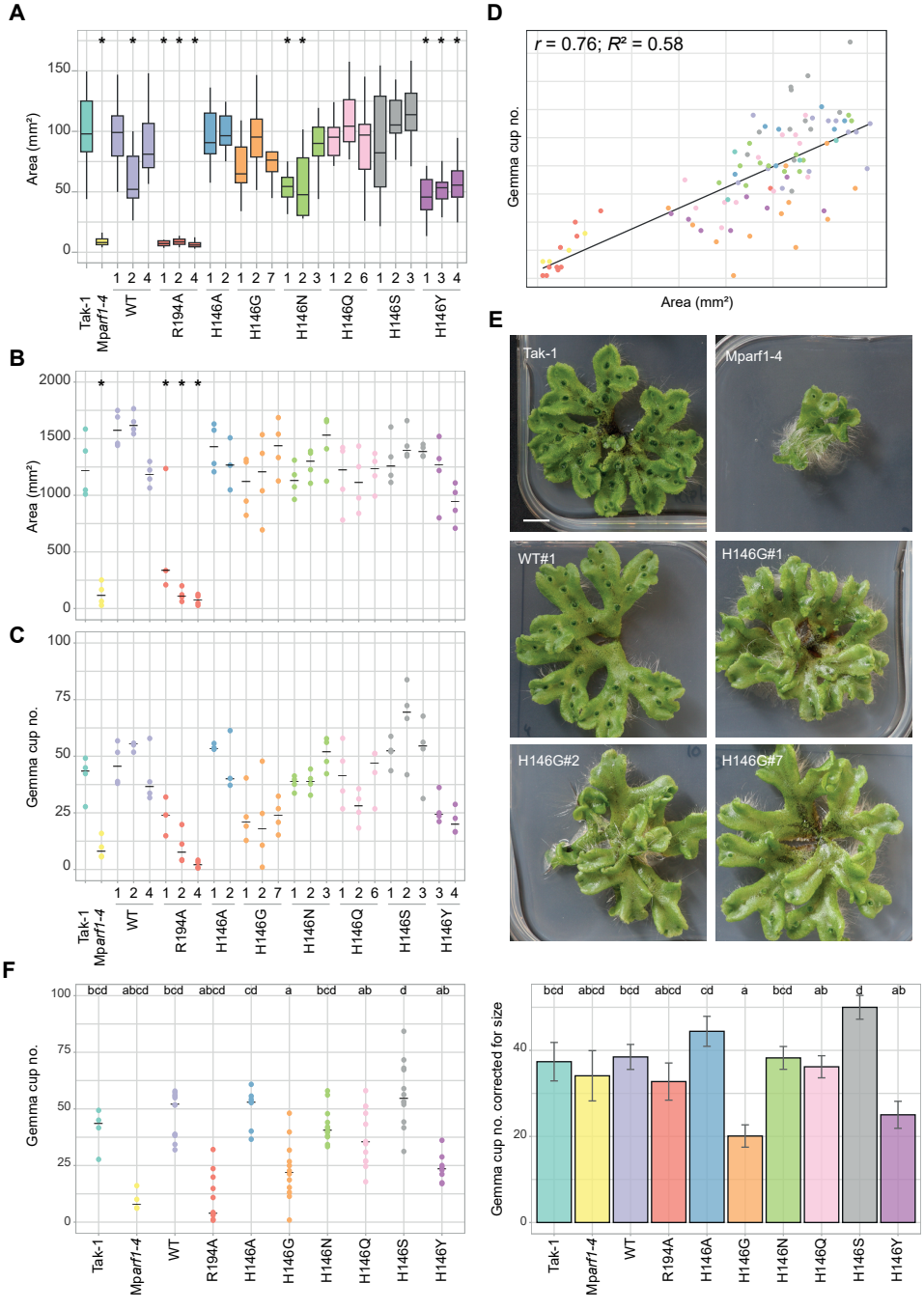

A
Area (mm²)
150
100
50
0
1 2 4 1 2 4 1 2 1 2 7 1 2 3 1 2 6 1 2 3 1 3 4
Tak-1
Mparf1-4
WT
R194A
H146A
H146G
H146N
H146Q
H146S
H146Y
B
Area (mm²)
2000
1500
1000
500
0
C
Gemma cup no.
100
75
50
25
0
1 2 4 1 2 4 1 2 1 2 7 1 2 3 1 2 6 1 2 3 3 4
D
r = 0.76; R² = 0.58
Gemma cup no.
Area (mm²)
E
Tak-1
Mparf1-4
WT#1
H146G#1
H146G#2
H146G#7
F
Gemma cup no.
bcd abcd bcd abcd cd a bcd ab d ab
Gemma cup no. corrected for size
40
20
0

Fig. 6 | Phenotyping of Mp*ARF1-H146* mutants. (A) Quantification of projected area from 14-day old gemmalings of Tak-1, Mp*arf1-4* and Mp*ARF1-H146* mutants, with $n \geq 15$ per independent transgenic line per experiment. **(B, C, D, E)** Quantification of projected area (B), number of gemma cups (C), scatterplot and Pearson correlation with projected area from B and gamma cup number from C (D), and pictures (E) of five week old gemmalings from same transgenic lines as in A. **(F)** Gemma cup numbers from C, pooled per genotype as observed (upper panel) and after correction for projected area by analysis of covariance (ANCOVA; lower panel). Asterisks in panels A, B and C, and letters in E represent statistical significant differences tested by ANOVA (A, B), or ANCOVA using projected area as covariate (C, E), both followed by post-hoc testing with Sidak correction ($p \leq 0.05$). Asterisks, rather than letters, were used for brevity in A to only show lines different from Tak-1, *WT#1* and *WT#4*. Scale bar in D is 10 mm.

and after which their projected area **(Fig. 6B)** and their gemma cup number **(Fig. 6C)** were measured. After 5 weeks, we noted all mutants except Mp*arf1-4* and the *R194A* lines to have similar projected area, in contrast to the smaller area observed after 14 days. The *H146Y* and *H146N* mutants therefore likely have reduced growth speed and eventually catch up in size. We reasoned that larger thalli might produce more gemma cups, which was confirmed by a moderate Pearson correlation ($r = 0.76$; **Fig. 6D**). We therefore performed an analysis of covariance (ANCOVA) on gemma cup numbers using projected size as a covariate, of which the F-test indicate differences between individual lines exist: $F(23, 69) = 14.5$, $p < 0.001$. Indeed, *H146G* lines visibly showed reduced gemma cup production **(Fig. 6E)**, although no individual line of **Fig. 6C** could be shown to have a different number of gemma cups than Tak-1 or the WT lines after post-hoc testing. Since we suspected a lack of statistical power in the post-hoc testing, we pooled all observations and performed an ANCOVA on each genotype – $F(9, 83) = 9.4$, $p \leq 0.001$ – rather than individual lines **(Fig. 6F)**. After post-hoc testing, genotype *H146G* indeed has reduced gemma cup production compared to Tak-1 or *WT* lines of equal thallus area. Since *H146G* lines were not affected in projected thallus area **(Fig. 6A, 6B)**, it suggests the *H146G* variant specifically alters gemma cup production.

All tested *MpARF1-H146* substitutions could fully or partially complement the growth defects of Mp*arf1-4*. This extended to complementing organogenesis of the antheridiophore **(Fig. S2)**, since at least one line per *H146* variant was able to produce antheridiophores in contrast to Mp*arf1-4*. These results suggest the *H146* variants should have a (partially) restored NAP. To test this, we grew all the lines on medium containing 3 µM NAA **(Fig. 7)**. First, all lines of *H146A*, *H146G* and *H146S* had a restored NAP, as indicated by NAA inhibiting their growth and triggering an overproliferation of rhizoids. The *H146A* and *H146S* variants can therefore complement Mp*arf1-4* on all tested phenotypes, and *H146G* complementing all phenotypes except gemma cup production. Second, two of the *H146N* lines (*H146N#1* and *H146N#2*) and all three *H146Y* lines showed partial complementation of thallus area after 14 days **(Fig. 6A)**, but otherwise normal

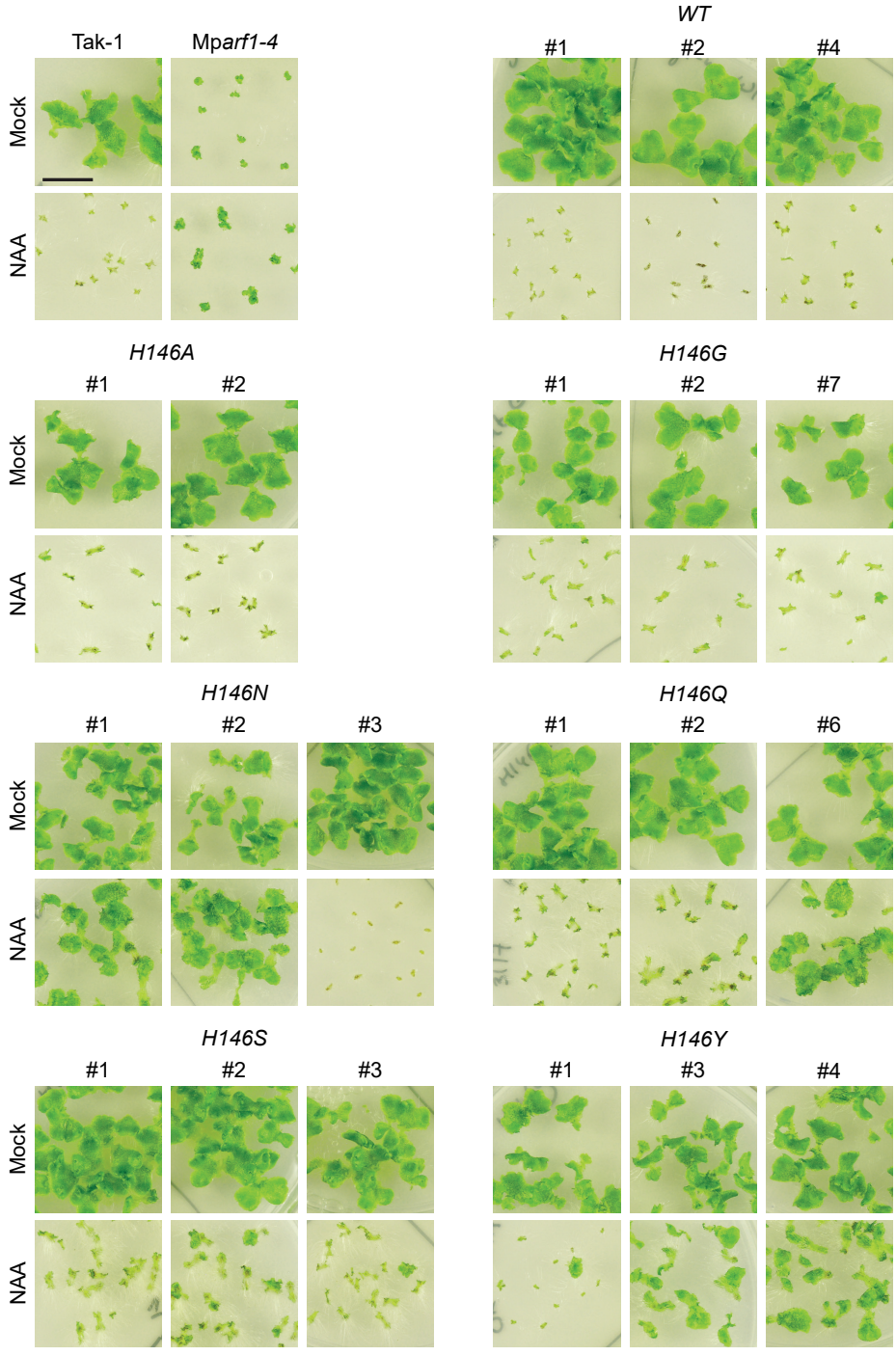


Fig. 7 | Mp*ARF1-H146X* mutants on NAA. Fourteen-day old gemmalings from *H146X* mutants, following naming scheme of Fig. 1. Gemmae were grown on medium without (mock) or with 3 µM NAA. Scale bar is 10 mm.

gemma cup numbers and restored antheridiophore production **(Fig. 6F, S2)**. Four of the five lines (except *H146Y#1*) were NAA resistant **(Fig. 7)**. In contrast, *H146N#3* fully complemented thallus area after 14 days and has a restored NAP. These results indicate that while *H146N* and *H146Y* generally have reduced growth speed, differences between individual lines exist.

The previous phenotyping experiments indicated that MpARF1-H146 can tolerate different mutations. To further test whether physiochemical properties cause *H146* variants to partially complement, we tested *H146E* and *H146F* mutants. Glu (E) has the same length as Gln (Q), and *H146Q* lines were able to fully complement most Mp*arf1-4* phenotypes **(Fig. 6, 7)**. However, while Gln is polar, Glu is negatively charged. Although Glu variants were found in A-ARFs **(Fig. 1B)**, Mp*arf1-4* complemented with Mp*ARF1-H146E* did not complement **(Fig. 8)**, indicating a negative charge is detrimental for ARF DNA-binding. *H146Y* lines were partially complementing, despite being a bulky hydrophobic residue. Tyr is composed of a hydrophobic benzyl ring and a polar hydroxyl group, the latter of which could hydrogen bond with the DNA. Phenylalanine (F) has the same benzyl group, but lacks the hydroxyl group. Nevertheless, we found Mp*arf1-4* transformed with Mp*ARF1-H146F* to partially complement **(Fig. 8)**, indicating even a fully hydrophobic residue could partially replace H146.

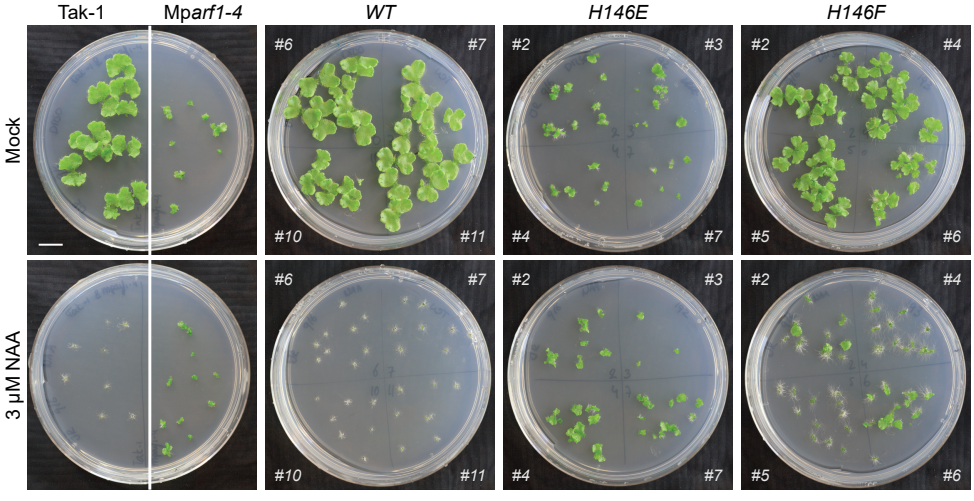


Fig. 8 | Mp*ARF1-H146E* and *H146F* mutants on NAA. Fourteen-day old gemmalings without (mock) and with 3 µM NAA. Shown are four independent lines, out of eight independent lines for Mp*ARF1-H146E* and *H146F* mutants.

H146N and Y mutants show reduced auxin sensitivity

Lines *H146N#1*, *H146N#2* and all three *H146Y* lines showed partial growth complementation after 14 days and auxin resistance, whereas line *H146N#3* fully complemented **(Fig. 6A, 7)**. Although each Mp*ARF1* variant is driven by $_{pro}$Mp*ARF1*, random T-DNA integrations could nevertheless cause differences in expression. Since the Mp*ARF1-H146* constructs were not Citrine-fused (see material and methods), we performed real-time quantitative PCR (RT-qPCR) to measure Mp*ARF1* expression in 10 day old gemmalings **(Fig. 9A)**. Both *WT#1* and *WT#4* were expressing Mp*ARF1-WT* at levels similar to Tak-1, whereas *WT#2* has 8-fold (3 $\log_2$) higher expression. While *WT#1* and *WT#4* also had comparable thallus area as Tak-1 after 14 days **(Fig. 6A)**, *WT#2* had a reduced thallus area. Reduced thallus

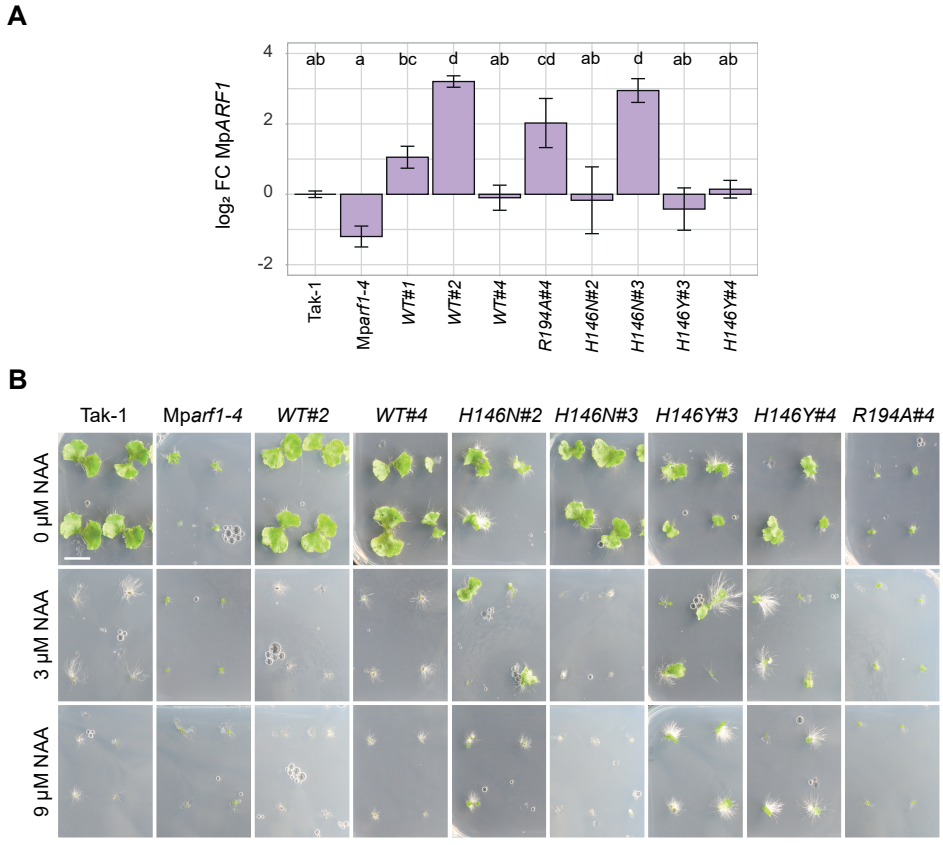


Fig. 9 | Auxin sensitivity and Mp*ARF1* expression of Mp*ARF1-H146N* and *H146Y* mutants. (A) Relative Mp*ARF1* expression shown as $\log_2$ fold change in fourteen-day old gemmalings measured by RT-qPCR. Expression was normalized by the geometric mean of Mp*APT3*, Mp*ACT7* and Mp*SAND*. Letters represent statistical significant differences tested by ANOVA and post-hoc testing with Sidak correction ($p < 0.05$). **(B)** Fourteen day old gemmalings of Tak-1, Mp*arf1-4* and specific lines of Mp*arf1-4* transformed with Mp*ARF1* with indicated variants grown on 0, 3 and 9 µM NAA. Scale bar is 10 mm.

area was previously observed for Mp*ARF1* overexpression lines ($_{pro}$Mp*EF1α:*Mp*ARF1*), which was also accompanied by higher auxin sensitivity (Flores-Sandoval *et al.*, 2015). For the partially complementing lines *H146N#2*, *H146Y#3* and *H146Y#4* we observed levels of Mp*ARF1* expression comparable to Tak-1, *WT#1* and *WT#4*, while the fully complementing line *H146N#3* has expression levels comparable to *WT#2* **(Fig. 9A)**. In other words, Mp*ARF1* expression levels influence the gemmaling growth speed.

Since Mp*ARF1* overexpression lines have a higher auxin sensitivity (Flores-Sandoval *et al.*, 2015), we tested whether the auxin-resistant phenotype of the *H146N* and *H146Y* lines **(Fig. 7)** was true resistance or lower sensitivity. To that end, we grew the *H146N* and *H146Y* lines on 3 and 9 µM NAA **(Fig. 9B)**. Both Tak-1 and the *WT* lines exhibited an auxin response at both 3 and 9 µM NAA while Mp*arf1-4* and line *R194A#*4 were unresponsive **(Fig. 9B)**. Line *H146N#3*, which showed full growth complementation but 8-fold higher Mp*ARF1* expression, responds to both 3 and 9 µM NAA, mirroring Tak-1 and *WT*. In contrast, while the partially complementing lines *H146N#2*, *H146Y#3* and *H146Y#4* did not respond to 3 µM NAA, they did respond at 9 µM NAA. These results indicate that *H146N#2* and the *H146Y* lines, but not *H146N#3*, have reduced auxin sensitivity, rather than full auxin resistance. Combined, these results together with the Mp*ARF1* expression indicate the NAP in these lines is restored, but the stoichiometry is different compared to the Tak-1 situation. In the case of *H146N#3*, higher expression of Mp*ARF1-H146N* can fully complement growth and the auxin response, which could be explained by higher expression compensating for weaker DNA-binding of Mp*ARF1-H146N*. In contrast, for the other *H146N* and *H146Y* lines, similar expression of MpARF1-H146N and MpARF1-H146Y as Tak-1 results in partial complementation, since reduced DNA-binding would results in a weaker auxin response.

The results above suggested that MpARF1-H146N and MpARF1-H146Y have lower DNA-binding affinity. However, due to MpARF1-H146's role in defining medium- versus high-affinity sites, it is also possible that MpARF1-H146N and MpARF1-H146Y regulate a different set of target genes, either lacking some target genes or binding novel ones. To address this, we performed RNA-seq on Tak-1, Mp*arf1-4*, *WT#4*, *H146N#2* and *H146Y#4*. The latter three transgenic lines were chosen for their similar levels of Mp*ARF1* expression to Tak-1 **(Fig. 9A)**. Previous transcriptomics with auxin on *M. polymorpha* has shown that the amplitude (i.e., maximum positive and negative fold-changes) of auxin-induced gene induction and repression is weak. For instance, a 1-hour treatment of Tak-1 gemmae with the synthetic auxin 2,4-dichlorophenoxyacetic acid (2,4-D) resulted in amplitudes less than $\log_2$ of 1 (Mutte *et al.*, 2018). These low amplitudes were mirrored in other transcriptomes (see Chapter 4, Fig 3). We reasoned that an increased amplitude could be achieved by first inhibiting MpARF1 by the accumulation of MpIAA before IAA treatment, as this would increase the range of change in gene expression. We therefore pre-treated gemmae with the auxin biosynthesis inhibitors L-kynurenine

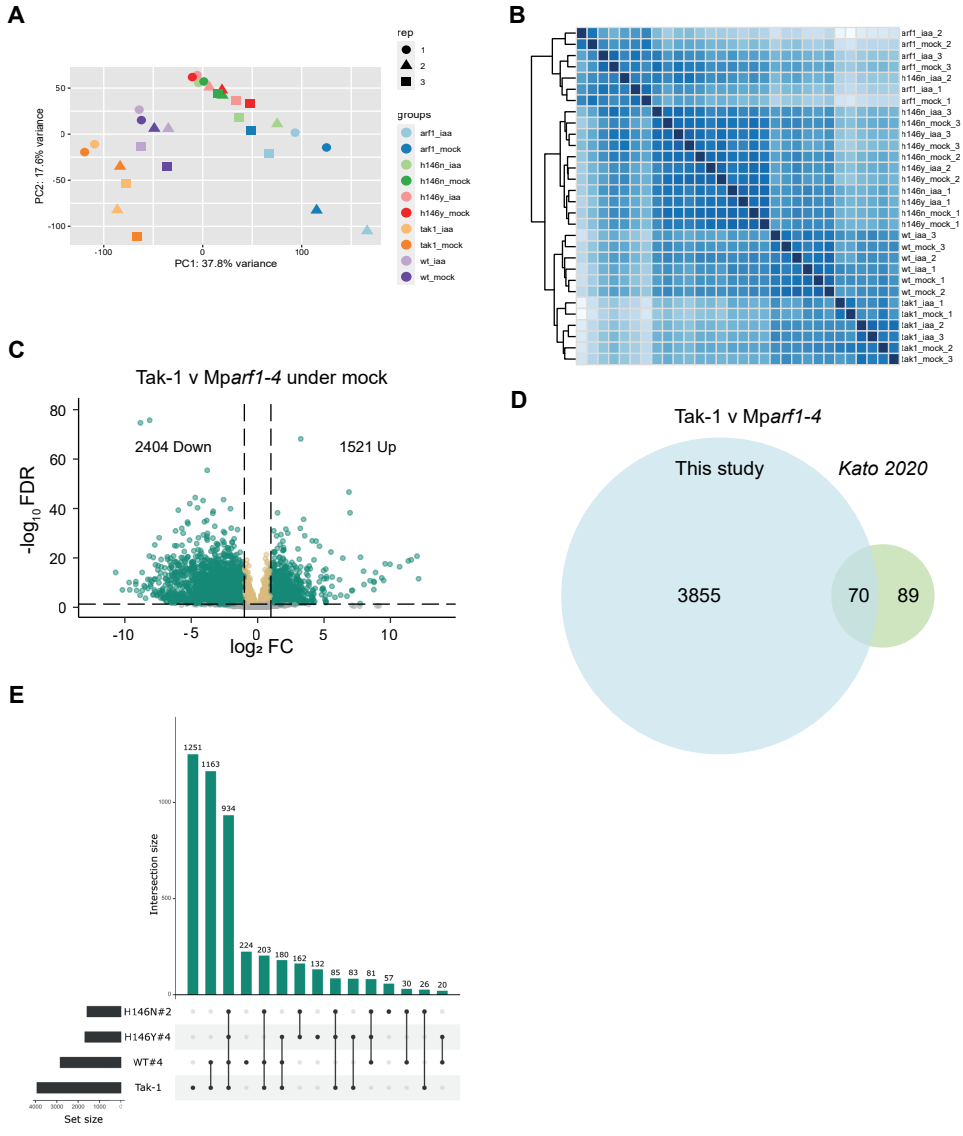


Fig. 10 | Quality measures of RNA-seq analysis. (A) PCA plot RNA-seq transcriptomes of Tak-1, Mp*arf1-4* and complementation lines *WT#4*, *H146N#2* and *H146Y#4* from three independent pools of >100 gemmae per pool, treated without (mock) and with 1 µM IAA. **(B)** Hierarchical clustering of RNA-seq transcriptomes. **(C)** Volcano plot plotting differentially expressed genes (DEGs) in Tak-1 versus Mp*arf1-4* under mock conditions. Green dots represent genes passing the $\log_2$FC (-1 ≤ log2FC ≥ 1) and false-discovery rate (FDR ≤ 0.05) threshold, beige dots passing only FDR threshold and grey dots passing neither threshold. **(D)** Tak-1 versus Mp*arf1-4* DEGs from this study compared to DEGs found for the same comparison in Kato *et al.*, (2020). **(E)** Venn diagram of number of DEGs from Tak-1, *WT#4*, *H146N#2* and *H146Y#4* versus Mp*arf1-4*.

(L-kyn) and Yucasin (Yuc) for 6 hours, a timing that was previously shown to induce maximum levels of endogenous MpIAA accumulation in gemmae (Das, 2023), followed by a 1-hour IAA or mock treatment. After the RNA-seq, the PCA plot and hierarchical clustering indicated genotypes and replicates, rather than IAA and mock treatments, clustered together **(Fig. 10A, 10B)**. Both methods indicate Tak-1 and Mp*arf1-4* form the two extremes, with *WT#1* clustering closer to Tak-1 and *H146N#2* and *H146Y#4* clustering in-between **(Fig. 10A)**. Indeed, hierarchical clustering suggests that both *H146N#2* and *H146Y#4* are closer to Tak-1 than Mp*arf1-4* **(Fig. 10B)**, consistent with their partially complementing phenotype. However, both clustering methods suggest the effect of IAA on the transcriptome is confounded by genotype effects. We reasoned this could be explained by the L-Kyn and Yuc treatments, since a six-hour treatment confounds the effects of different MpARF1 variants. For instance, comparing Tak-1 with Mp*arf1-4* under mock conditions, which would show the effects of L-Kyn and Yuc treatment, we found a total of 3952 differentially expressed genes (DEGs; **Fig. 10C**), which is in stark contrast to the 161 DEGs that were found previously **(Fig. 10D)** (Kato *et al.*, 2020). Indeed, when comparing all genotypes to Mp*arf1-4* under mock conditions, we found hundreds of genes unique to each genotype or subset of genotypes **(Fig. 10E)**. Nevertheless, despite the confounding effects, the main MpARF1 targets should still respond to IAA, and we therefore focused on mock versus IAA comparisons.

When comparing IAA and mock treatments for Tak-1, 112 genes were differentially regulated: 43 downregulated and 69 upregulated **(Fig. 11A)**. In contrast, no genes were auxin-regulated in Mp*arf1-4*, as discussed previously (Chapter 4). In the other genotypes, we found only a limited number of auxin-regulated DEGs in *WT#4* (19 up, 12 down) and only 5 and 8 DEGs total for *H146N#2* and *H146Y#4* respectively **(Fig. 11B)**. With lower auxin sensitivity for *H146N#2* and *H146Y#4* **(Fig. 9)** it could be that auxin-regulated DEGs in these lines fall under the used fold-change threshold ($-1 \leq \log_2 FC \geq 1$). We therefore subset the Tak-1 IAA-mediated DEGs and performed Pearson correlation with the IAA-mediated fold changes of the same genes in the other genotypes **(Fig. 11C)**. There was no correlation between Tak-1 and Mp*arf1-4* ($r = 0.18$) and fold changes of these genes were small ($\log_2 FC$ of -0.29 and 0.18 on average) in Mp*arf1-4*, further corroborating that Mp*arf1-4* does not respond to auxin. In contrast, *WT#4* has a high correlation with Tak-1 ($r = 0.92$), although the amplitude for the upregulated genes is smaller in *WT#4* ($\log_2 FC$ between -3.3 to 3.7) compared to Tak-1 ($\log_2 FC$ between -3.3 to 9.8), indicating the auxin response is restored but slightly weaker in *WT#4*. For *H146N#2* and *H146Y#4*, we found high correlations with Tak-1 ($r = 0.80$ and 0.72 respectively). This high correlation indicates that the majority of Tak-1 IAA-mediated genes are regulated similarly in the *H146* variants, but with a strongly reduced amplitude. In conclusion, while we were unable to determine if MpARF1-H146N and MpARF1-H146Y bind a novel set of target genes, the results support that MpARF1-H146N and MpARF1-H146Y have a weaker auxin response in general, which is possibly caused by lower DNA-binding affinity.

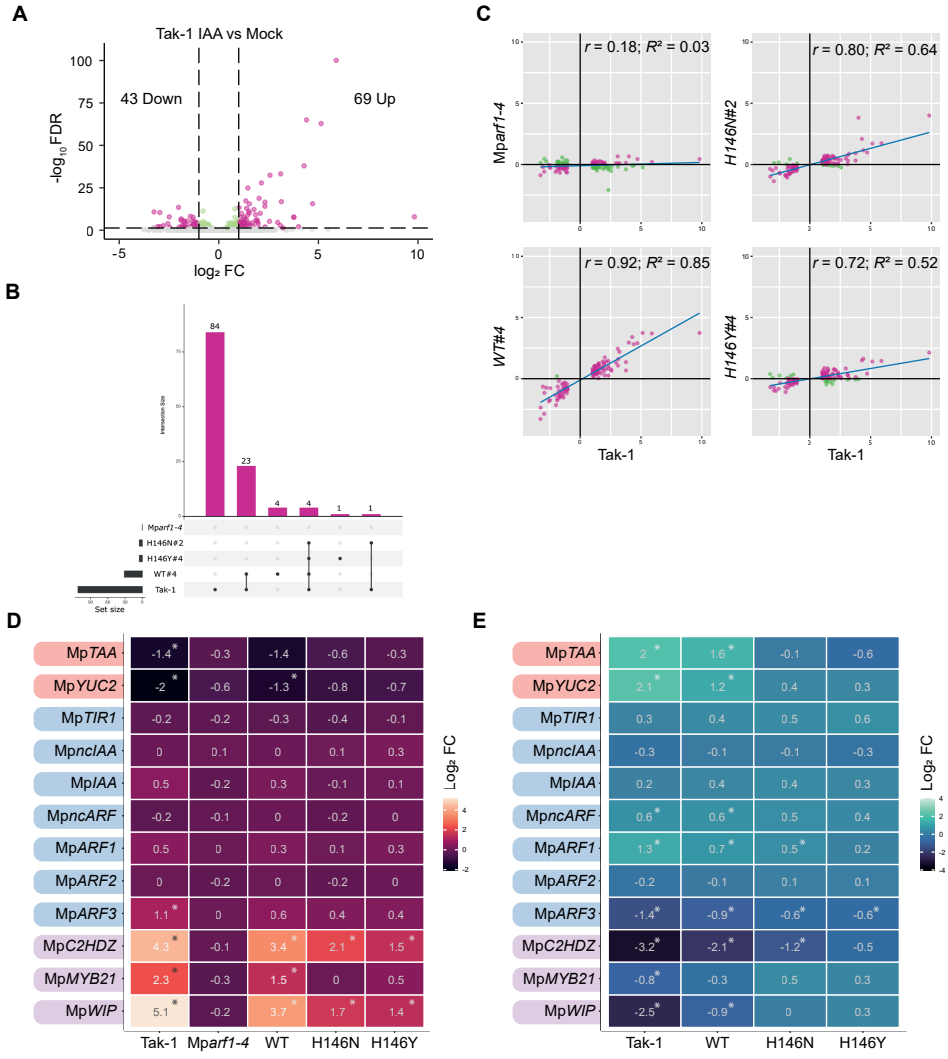


Fig. 11 | RNA-seq analysis. (A) Volcano plot showing 112 DEGs in Tak-1 comparing IAA versus mock treated samples. Magenta dots represent genes passing the Log_2 FC ($-1 \leq Log_2 FC \geq 1$) and FDR (FDR ≤ 0.05) threshold, green dots passing only FDR threshold and grey dots passing neither threshold. **(B)** Upset plot showing IAA versus mock DEGs in all genotypes. **(C)** Scatterplots and Pearson correlations (blue lines) of Tak-1 IAA versus mock DEGs of A (x-axis) and their corresponding Log_2 FC in mutant lines (y-axis). Dots represents genes following the same pattern (magenta) or the opposite pattern (green) in both samples. **(D, E)** Heatmaps of Log_2 FC values for selected genes, comparing IAA versus mock (D) and genotype versus Mp*arf1-4* (E). Color scale and values within each cell indicate Log_2 FC, asterisks indicate statistical significance (FDR ≤ 0.05). Balloons around the gene ID indicate if a gene is related to auxin biosynthesis (red), a NAP component (blue), or an MpARF1 target gene (purple). Note Mp*R2R3-MYB21* is abbreviated to Mp*MYB21*.

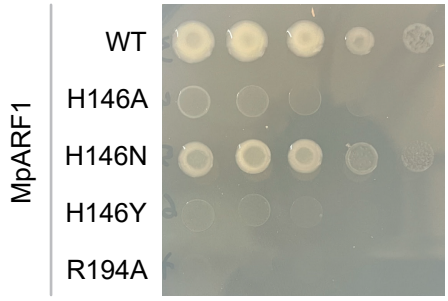
While the above results explain why *H146N#2, H146Y#3* and *H146Y#4* can partially complement Mp*arf1-4* **(Fig. 6A)**, i.e., by binding similar target genes as MpARF1-WT, it does not explain why the auxin-regulated amplitude of the target genes is so weak in the mutants **(Fig. 11C)**, nor why these mutants exhibit reduced auxin sensitivity **(Fig. 10A)**. One explanation could be a lack of MpARF1-regulated feedback loops with other auxin-related genes in the *H146* variants. Since previous studies indicate that auxin application triggers an MpARF1-mediated negative feedback loop to auxin biosynthesis (Eklund *et al.*, 2015; Mutte *et al.*, 2018), we investigated this possibility by plotting the fold changes of auxin-related genes (see Supp. Table 1) comparing mock and IAA treatments **(Fig. 11D)**, and comparing all genotypes to Mp*arf1-4* under mock conditions, which would show the effects of L-Kyn and Yuc treatment and possible repression through MpARF1 **(Fig. 11E)**. When comparing Tak-1 and Mp*arf1-4*, we found that the known MpARF1-regulated downstream genes Mp*C2HDZ*, Mp*R2R3-MYB21* and Mp*WIP* (Kato *et al.*, 2015; Kato *et al.*, 2017; Mutte *et al.*, 2018; Suzuki *et al.*, 2023) were induced by IAA in Tak-1 and not in Mp*arf1-4* **(Fig. 11D)**. Under mock conditions, these genes are repressed in Tak-1 compared to Mp*arf1-4* **(Fig. 11E)**, thereby corroborating previous results (Kato *et al.*, 2015; Kato *et al.*, 2020) that MpARF1 switches between repressive and activating roles *in vivo* depending on the auxin concentration. Indeed, the majority of the 112 IAA-mediated DEGs in Tak-1 **(Fig. 11A)** show the opposite pattern when comparing Tak-1 versus Mp*arf1-4* under mock conditions (r = -0.70; **Fig. S3**). From the NAP genes, the auxin biosynthesis genes Mp*TAA* and *MpYUC2* were indeed downregulated in Tak-1 but not Mp*arf1-4* after IAA treatment **(Fig. 11D)**, corroborating evidence for an MpARF1-dependent feedback loop (Mutte *et al.*, 2018). Notably, Mp*TAA* and Mp*YUC2* are downregulated in mock conditions in Tak-1 compared to Mp*arf1-4* **(Fig. 11E)**, indicating that the feedback loop works both ways to maintain auxin homeostasis. Of the NAP components Mp*TIR1*, Mp*IAA*, Mp*ARF1* and Mp*ARF2*, none showed expression changes in IAA conditions **(Fig. 11D)**, and Mp*TIR1*, Mp*IAA* and Mp*ARF2* additionally have similar expression levels in each genotype compared to Mp*arf1-4* **(Fig. 11E)**. Only Mp*ARF1* is higher in each genotype compared to Mp*arf1-4*, which we also noted from RT-qPCR **(Fig. 9A)**. This indicates no feedback loops occur between MpARF1 and auxin signalling components, but instead, *M. polymorpha* readjusts auxin concentrations through an MpARF1-mediated negative feedback loop.

When comparing auxin-related gene expression profiles between Tak-1 and *WT#4*, *H146N#2* and *H146Y#4* under IAA conditions and compared to Mp*arf1-4* **(Fig. 11D, 11E)**, they mirror the overall trends of these mutants **(Fig. 11C)**, that is, both repression and activation of auxin biosynthesis genes and MpARF1 targets were weaker. Given that Tak-1 and *WT#4* exhibit a phenotypic auxin-response at 1 µM NAA **(Fig. 9B)**, the fold-changes between mock (i.e. effects of inhibition of auxin biosynthesis) and IAA conditions for Tak-1 **(Fig. 11A, 11D)** likely represent the full range of fold-changes. In the *H146N#2* and *H146Y#3* these same fold-changes are likely only reached at IAA/NAA concentrations higher than 3 µM **(Fig. 9B)**. These results are therefore consistent with

lower DNA-binding affinity of MpARF1-H146N and MpARF1-H146Y.

H146N and H146Y have reduced and altered DNA-binding mechanisms

The results above for the *H146N#2* and *H146Y#4* mutant lines could be explained by lower DNA-binding affinity. We therefore used a bacterial 1 hybrid (B1H) setup to test the DNA-binding of *MpARF1-H146N* and *MpARF1-H146Y* to a *3xDR5*, TGTCGG motif **(Fig. 12)**. In this assay, the ability of ARFs to bind to their target sites translates to growth responase. We find that the Mp*ARF1-R194A* mutant was unable to grow due to a failure to bind the *3xDR5* motif. All three *H146* mutants have reduced bacterial growth compared to *WT*, thereby confirming the observed phenotypes and muted transcriptome results in *H146N* and *H146Y* lines are due to weaker DNA-binding affinity.



F**ig. 12 | Bacterial 1 hybrid (B1H) assay with MpARF1 variants.** B1H assay using a *3xDR5* motif as prey plasmid and Mp*ARF1-DBD* construct without (WT) or with indicated mutations as bait plasmid. From left to right are 10-fold dilutions (10^0 to 10^{-5}). Note that figure shown is the same as Chapter 3, Fig. 1. Assay was repeated twice with identical results.

To understand the atomic mechanisms why some H146 variants were capable to complement and others were not, we performed molecular docking **(Fig. 13)**. All interactions found from docking are schematically shown in **Fig. 13H**). Both H146A and H146S variants were able to fully complement the Mp*arf1-4* mutant **(Fig. 6, 7)**, although H146A appeared to have lower DNA-binding affinity **(Fig. 12)**. However, both Ala and Ser are small residues that did not form hydrophobic or hydrogen bonds with any of the nearby base pairs **(Fig. 13B, 13F)**. It is therefore likely these variants are not affected in specificity, but rather by affinity for all TGTCNN elements globally. In contrast, H146Q was also able to complement **(Fig. 6, 7)**, and we observed in docking that Q146 could rotate deeper into the major groove than H146 and hydrogen bond to G5 **(Fig. 13E)**, thereby still offering base specificity. The remaining variants, H146E, H146N and H146Y, can all dock to the TGTCGG element according to HADDOCK. For H146Y, while H146Y and H146F can partially complement **(Fig. 6, 8)** and can fit in the major groove **(Fig. 13G)**, it is surrounded by the polar groups from G6, C7:G7' and G8:C8'. Therefore, while MpARF1-H146Y must be able to enter the major groove, since otherwise *H146Y* and *H146F* were unable to complement at all, the hydrophobic residues make the interaction energetically less

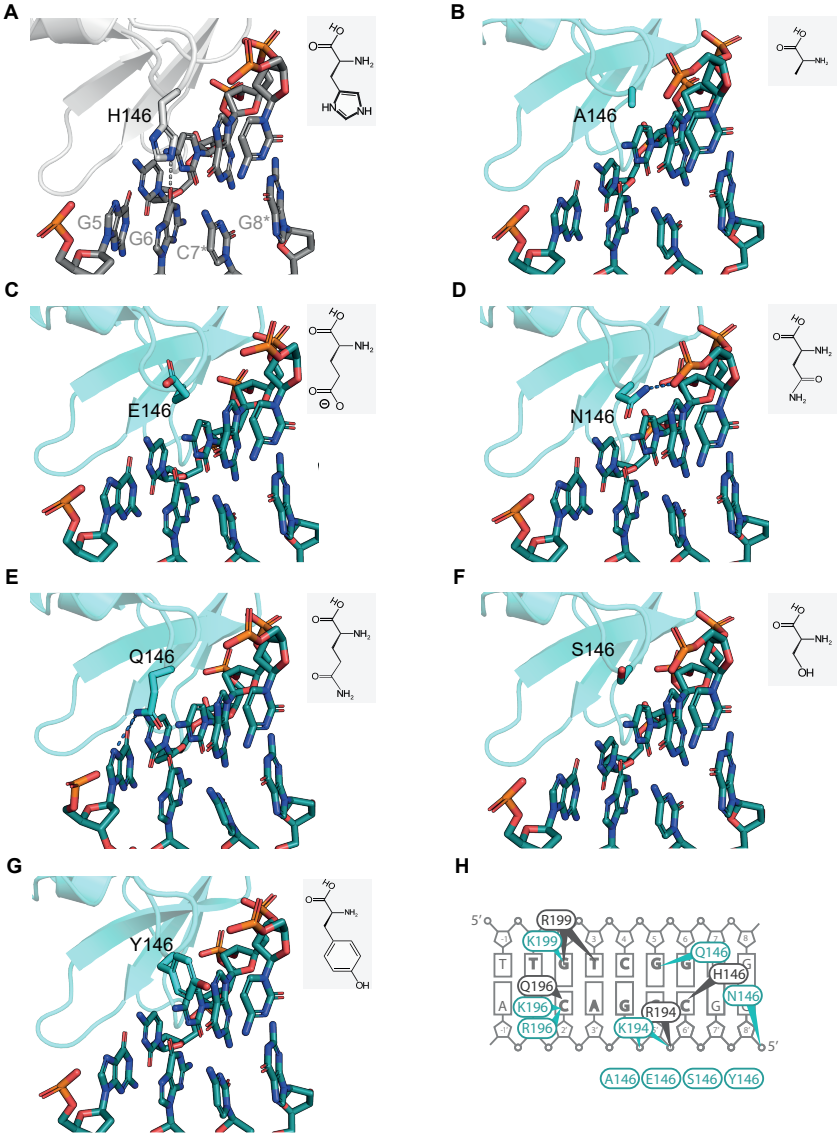


Fig. 13 | Molecular docking of MpARF1-H146X variants. (A-G) Each panel shows the ARF-AuxRE interface of wild type MpARF1 (crystal structure of MpARF2, referred as MpARF1) (A), H146A (B), H146E (C), H146N (D), H146Q (E), H146S (F) and H146Y (G). Colored sticks in A-G indicate either the wild type interface (white/grey) or docked interface (cyan/teal). Nucleotides are indicated in A and correspond with panels B-G. Asterisks (C7* and G8*) indicate nucleotide is outside of TGTCGG AuxRE. Dashed lines indicate hydrogen bonds in the wild type (white) and docked (marine) structures. Chemical structures of amino acids are shown next to each panel. **(H)** Schematic overview of DNA-interactions from wild type MpARF1 (grey) and after docking MpARF1-variants (cyan). Note A146, E146, S146 and Y146 do not form any hydrogen bonds in the docked structures.

favorable. Finally, both H146E and H146N were able to dock **(Fig. 13C, 13D)**. Their failure to complement (H146E) or partially complement (H146N) could be shown by considering their rotamers. For MpARF1-H146N, N146 bonded with the phosphate group of C8' in the docking structure, while an alternative rotamer could hydrogen bond with G7 **(Fig. 14A)**. Neither nucleotide is part of the TGTCNN element proper, and since MpARF1-WT does not show preferences for residues downstream of the TGTCNN element (Chapter 4) it is possible certain TGTCNN elements are excluded from being bound by H146N by having unfavorable nucleotides at positions 7 and 8. Finally, while Glu (E) and Gln (Q) have the same length of their side chain, Gln has both a hydrogen bond acceptor (NH) and donor (O) group on its head, which allows it to hydrogen bond to G6 **(Fig. 13E)**. In contrast, Glu only has two donor groups and is negatively charged, which explains why E146 is rotated away from the DNA in the docking structure **(Fig. 13C)**, since every group in its vicinity is either a negatively charged phosphate group or hydrogen bond donor **(Fig. 14B, 14C)**. Therefore, MpARF1-H146E is likely unable to bind DNA due to the electrostatic clashes.

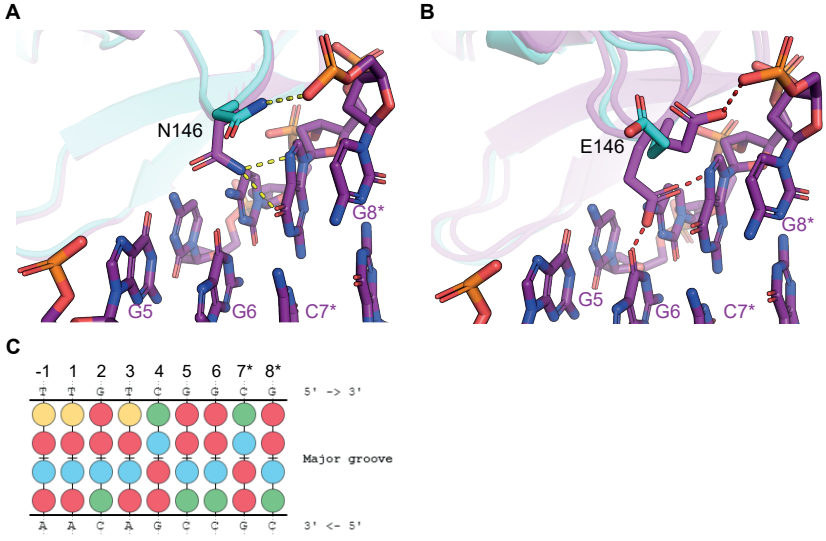


Fig. 14 | Alternative rotamers in docked structures of H146N and H146E. (A-B) Docked structures and alternative rotamers of H146N (A) and H146E (B). Original docked structures are same as Fig. 13 (cyan) and show overlayed alternative rotamers (purple). Nucleotides are indicated and asterisks (C7* and G8*) indicate nucleotides outside of TGTCGG AuxRE. Dashed lines indicate hydrogen bonds between hydrogen bond donor and acceptor (yellow) or electrostatic clashes between hydrogen bond acceptor and acceptor (red). **(C)** Overview of base readout in the major groove of the DNA molecule used in docking. Circles represent hydrogen bond acceptor (red), hydrogen bond donor (blue), methyl groups (yellow) and heterocyclic-hydrogen (green). Made with DNA-readout viewer (Adam *et al.*, 2019).

Discussion

The B3 domain of the ARFs was shown to be extremely conserved, and the DNA-binding residues in particular are even conserved in ARFs from algal species (Chapter 3). Not only did this suggest an evolutionary conserved DNA-binding mechanism for ARFs from all plant lineages, it also implies that mutations on DNA-binding residues are functionally detrimental and undergo purifying selection. In this chapter, we aimed to establish whether variation for DNA-binding residues exist in the ARF family, and what their effect is on DNA-binding.

We found that changes to the TGTC-core binding residues R194, P197 and R199 were detrimental for the functioning of MpARF1. Transcription factors and their DNA elements coevolve to retain the best fit (Garvie and Wolberger, 2001; Harteis and Schneider, 2014), which suggest that the tested MpARF1 variants of R194, P197 and R199 **(Fig. 1, 3)** are negative mutations that deviate from the best fit and cause loss of DNA-binding. In other words, it is unlikely that ARFs with the tested variants can bind TGTCNN-elements, unless variation on secondary residues can compensate for losing one of the primary DNA-binding residues. It will therefore be interesting to perform DNA-binding assays with these ARFs to see whether they are still capable of binding TGTCNN-elements. Most of the variation for R194, P197 and R199 was found in four ARFs from angiosperm C-ARFs. Since they have variation on all three residues, it is very likely that they cannot bind TGTCNN-elements. Whether this indicates that these ARFs have become subspecialized or pseudogenes is unknown and would require studying these ARFs in detail. Finally, although the tested variants likely lost binding to TGTCNN-elements, they may not have lost DNA-binding completely. For instance, AtARF5-R215A and P218A mutants (homologous to MpARF1-R194A and P197A) gained binding to non-TGTCNN-elements instead. Testing DNA-binding preferences of the MpARF1 variants could therefore indicate whether DNA-binding is lost, or whether specificity is altered.

The alanine scanning results for MpARF1-R194, P197 and R199 only partially match previous, similar experiments with the homolog AtARF5. In this chapter, none of the alanine mutants were able to rescue Mp*arf1-4* **(Fig. 1C)**. In Arabidopsis in contrast, the homologous AtARF5-R215A and R220A mutants could partially complement the *Atarf5* mutant *mp-B4919*, while the AtARF5-P218A mutant could fully complement (Boer *et al.*, 2014). Whether these discrepancies are caused by different mutants, (Boer *et al.*, 2014), or by intrinsic properties of AtARF5 is unknown. Since both MpARF1-R194A and AtARF5-R215A lose DNA-binding *in vitro* (**Fig. 12**; Boer, *et al.* 2014), it does suggest that the discrepancies are caused by the mutant backgrounds. Testing both MpARF1 and AtARF5 variants in the same *in vitro* DNA-binding assays could reveal whether the proteins have different tolerances to mutations of R194, P197 and R199.

Surprisingly, MpARF1-H146 and MpARF1-Q196 are very amendable to alternative residues. This contrasts with the strong conservation of these residues in the A-ARFs, both between *Marchantia* accessions as well as within the plant kingdom. For *AtARF5*, the *H170A* mutant only partially complemented *mp-B4919* (Boer *et al.*, 2014), whereas *MpARF1-H146A* complemented fully **(Fig. 6, 7, S2)**. This illustrates this residue could play an important role in conditions or tissues not tested here. For instance, His residues have been shown to act as pH-sensitive switches in other TFs, altering protein folding, dimerization or DNA-binding itself in a reversible, pH-dependent manner (Chun *et al.*, 2023; Deochand *et al.*, 2016; Kang *et al.*, 2018). Plants experience intracellular pH fluxes under stress, for instance when flooded (Lehmann *et al.*, 2021), that could thereby alter MpARF1 target genes in a pH-dependent manner. Further biochemical and *in vivo* studies with the Q196 and H146 variants generated here will further elucidate their roles.

We therefore found that ARF-AuxRE follows a best fit principle wherein DNA-binding has been evolutionary optimized. Mutations in the core TGTC-binding residues are detrimental to MpARF1 function which, explains their strong conservation in A, B and C-ARFs. Mutations on Q196 and H146 are well-tolerated, despite their evident strong conservation, and in the case of H146 can lead to specific phenotypic effects. The MpARF1 variants engineered in this chapter will be a valuable resource for further studying ARF-AuxRE binding preferences.

Material & Methods

Plasmid construction

Gateway entry vectors (pENTr221) containing full-length Mp*ARF1* coding sequence (M1-P925) with and without stop codons were kindly provided by Hirotaka Kato and renumbered for this study as *pJR84_ENTr_MpARF1Cds** and *pJR86_ENTr_MpARF1CdsNo**. Note vector *pJR86* was domesticated for Golden Gate cloning by removal of an internal *BsaI* site (MpARF1-C543T). Primers used in this study are listed in Supplementary table 2. Site-directed mutagenesis was performed with a split-vector amplification method. To create H146 mutations, *pJR84* or *pJR86* were amplified with primer combinations JR428-JR376, and JR429-JR375 using Q5 High Fidelity PCR (M0492, New England Biolabs) followed by *DpnI* digestion (FD1703, Thermo Scientific). PCR amplicons JR428-JR376 and JR429-JR375 were fused with oligo MpARF1N_mut, which contained the required codon change, using HiFi Assembly (E2621, New England Biolabs). R194, Q196, P197 and R199 mutations were created similarly, amplifying *pJR84* and *pJR86* with JR428-JR378 and JR429-JR377 and using oligo MpARF1C_mut with mutated codons. Gateway destination vectors *pHKDW038* ($_{pro}$Mp*ARF1:LR-Citrine*) and *pHKDW046* ($_{pro}$Mp*ARF1:LR*) were kindly provided by Hirotaka Kato and used previously (Kato *et al.*, 2017; Kato *et al.*, 2020), and both vectors contain the 4869 bp upstream of the TSS of Mp*ARF1* (Mp1g12750) as $_{pro}$Mp*ARF1*. Mp*ARF1* were subcloned from *pJR84* to *pHKDW046* to create *pJR66_proMpARF1:MpARF1* and *pJR86* to *pHKDW038* to create *pJR89_proMpARF1:MpARF1-Citrine*, using LR clonase II (11791020, Invitrogen). The following mutants were cloned with *pJR66* as base: *H146A, H146G, H146N, H146Q, H146S, H146Y* and *R194A*. The following were cloned with *pJR89* as base: *H146E, H146F, R194K, Q196A, Q196K, Q196R, P197A, P197Q, P197R, R199A, R199K* and *R199E*.

Plant materials and culture conditions

The male accession of *M. polymorpha*, Takaragaike-1 (Tak-1), the null mutant Mp*arf1-4* in Tak-1 background (Kato *et al.*, 2017) and generated mutant lines were maintained axenic and asexually. Gemmae or sectioned apical notches (~ 10 mm^2) were propagated on half-strength Gamborg's B5 medium (hereafter ½ B5; G0209, Duchefa Biochemie) containing 1% agar (hereafter solid ½ B5) under 50-60 µmol photons m^{-2} s^{-1} white fluorescent light at 22°C under long day (16h light, 8h dark) conditions.

Transformation of Mp*arf1-4*

Mp*arf1-4* thallus material was transformed with *pJR66* or *pJR89* using an *Agrobacterium tumefaciens* protocol modified from (Kubota *et al.*, 2013). Mp*arf1-4* thallus material was grown anywhere between 2 weeks and 3 months (before the onset of dying tissue) prior to transformation. Green thallus material was cut indiscriminately into tiny pieces (< 1

mm^2) using a broad razor blade and added into 19 ml of liquid 0M51C medium. To this, a 3 ml overnight culture of *A. tumefaciens*, pelleted and resuspended in 1 ml 0M51C, was added and incubated for 1-2 days at 22°C under long day conditions. Thalli pieces were washed by centrifuging and decanting the 0M51C medium, followed by three washing steps with 30 ml of sterile water with 0.005% Triton-X100. At each washing step, thalli pieces were vortexed thoroughly for 5 min. After the last wash, thalli pieces were incubated in 10 ml water with 0.005% Triton-X100 and 1 mg/ml cefotaxime for 1-2 hours on an orbital shaker. Thalli pieces were grown on solid ½ B5, with 100 µg/ml cefotaxime, 10 mg/ml hygromycin (Mp*arf1-4* background) and 0.5 µM chlorsulfuron (*pJR66* and *pJR89*). Non-chimeric lines (G_1 generation) were established from T_1 lines via single gemmae. Gemmae from G_1 plants (G_2 generation) or later generations, from multiple independent T_1 lines were used for experiments.

Phenotyping experiments

Fourteen day growth experiments with or without NAA were performed by growing gemmae on solid ½ B5 with 3 µM NAA or without (adding an equivalent volume of DMSO) and growing at 50-60 µmol photons m^{-2} s^{-1} continues white fluorescent light at 22°C. In the case of the 3 and 9 µM NAA experiment, the volume of DMSO per plate was kept equivalent by diluting a 9 mM NAA stock to 3 mM with DMSO and an equivalent volume of DMSO was added to mock. After 14 days, pictures were taken with a Canon EOS250D camera with EF-S 18-55 IS STM lens. To quantify thallus growth area, pictures were resized to 2000x2000 pixels with ImageMagick (version 7.1.1-21), followed by pixel classification and object quantification of projected area of single gemmalings with Ilastik (version 1.3.3post3; Berg *et al.*, 2019).

For gemma cup number and 5 week growth experiment, gemmae were grown under the same conditions as above, except for 5 weeks. Projected thallus area was measured as above and gemma cup number counted manually.

To score antheridiophore production, a single gemmae per line was grown for 14 days at the conditions above, after which they were supplemented with far-red (FR) light (50-60 µmol photons m^{-2} s^{-1}). Pictures were taken after five weeks at FR conditions.

Transcriptome analysis

Plant material and growth conditions

Plants used for RNA-seq were grown as follows. Gemmae of Tak-1, Mp*arf1-4*, *WT#4*, *H146N#2* and *H146Y#4* were grown on solid ½ B5 with 2% sucrose following above described culturing conditions. Sucrose was added to enhance gemma cup production (Ishizaki *et al.*, 2016). After 5-6 weeks of growth, gemmae were harvested by holding the plate upside down above a tray filled with sterile demineralized water and hammering

the bottom of the plate with a metal spoon. Gemmae that fell in the water were collected in a 70 µm cell strainer. Gemmae were immediately submerged in a solution of liquid ½ B5 with 50 µM each of L-Kyn and Yuc. After 6 hours, gemmae were collected in a strainer and washed three times with sterile demineralized water. Gemmae were either submerged into liquid ½ B5 (mock) or liquid ½ B5 with 1 µM IAA (mock contained equivalent volume of DMSO). After 1 hour, gemmae were collected in 2 ml Eppendorf tubes and frozen in liquid nitrogen. Samples were kept at -70°C until processing. Replicates 1, 2 and 3 were collected in 1 week intervals.

RNA extraction and sequencing

Frozen gemmea were ground into a powder with a bead shaker. RNA was extracted with the RNeasy Plant Mini Kit (74904, Qiagen) including an on-column DNase treatment (69104, Qiagen). RNA-seq library construction was performed with the TruSeq kit (Illumina) and 150-bp paired-end sequencing with BGISEQ-500 was performed by BGI Tech Solutions (Hong Kong). Obtained raw fastq reads were quality controlled using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Cleaned reads were mapped to the *M. polymorpha* genome (v6.1, https::/marchantia.info/) using Hisat2 for paired-end reads (Kim *et al.*, 2019), followed by counting reads per gene with FeatureCounts using default parameters (Liao *et al.*, 2014). Finally, genes with low counts (<90 across all samples together) were removed and differentially expressed genes were calculated with DESeq2 (Love *et al.*, 2014).

RT-qPCR

Ten-day old gemmalings of each line used for RT-qPCR **(Fig. 9A)** were harvested and immediately frozen in liquid nitrogen. After being ground to powder in a bead shaker, RNA was extracted as described above, and 1 µg total RNA was subjected to cDNA synthesis (iScript cDNA Synthesis Kit; 1708891, Bio-Rad) according to the manufacturer's instructions. RNA was quality checked on nanadrop and gel electrophoresis and absence of genomic DNA was confirmed with a -RT control.

RT-qPCR was performed using a 384CFX Connect Real-Time PCR Detection system (Bio-Rad) and iQ SYBR Green Supermix (1708880, Bio-Rad). A two-step cycle of 95°C 10s followed by 60°C for 30s was repeated for 40 cycles, followed by a melt-curve analysis. Three biological and two technical replicates were used. All primers used are listed in Supplementary table 2. The geometric mean of Mp*SAND*, Mp*APT7* and Mp*APT3* (Saint-Marcoux *et al.*, 2015) was used to normalize expression of Mp*ARF1* according to (Taylor *et al.*, 2019).

Expression of MpARF1 variant proteins in live gemmae

MpARF1-Citrine fused proteins were detected using live gemmae imaging, which was

performed with a Leica SP8X-SMD confocal microscope equipped with a hybrid detector and a pulsed white-light laser (WLL). mCitrine was excited at 514nm using a notched filter laser line with a scan speed of 400 Hz (at 6% laser output of WLL). Fluorescence emission was detected between 521 and 564 nm, 2 linse and 6 frames were accumulated respectively. All images were acquired with a water immersive 20X objective lens. Fluoresence was detected in photon counting mode and time-gated detection was used to suppress chlorophyll autofluorescence.

Bacterial 1 hybrid assay

Both cloning of the bait and prey vectors and the general procedure of B1H were described before (Chapter 4). Assay was performed two independent times with new transformed *E.coli* with similar results.

Molecular Docking and structural analysis

Molecular docking of DNA-binding variants was performed using HADDOCK (Dominguez *et al.*, 2003; van Dijk *et al.*, 2013; van Zundert *et al.*, 2016). The MpARF2 structure (PDB: 6sdg) was used as template. The MpARF2 dimer (chain A and B) and DNA (Chain C and D) were written as separate PDB files. Chains B and D were renumbered (B +400, D +30) to conform with the unique identifier requirement of HADDOCK. Mutants in DNA-binding residues were made in both chains using the *mutation* tool of PyMOL Molecular Graphics System (version 2.5.5, Schrödinger, LLC), followed by energy minimization of the entire dimer in Swiss PDB Viewer (http://www.expasy.org/spdbv/; Guex and Peitsch, 1997).

HADDOCK was performed in 'easy' mode with default settings. Active DNA residues were: DNA: 1, 2, 3, 4, 5, 6, 14, 15, 16, 17, 18, 19, 31, 32, 33, 34, 35, 36, 44, 45, 46, 47, 48, 49; MpARF2: 136, 141, 145, 148, 186, 189, 191, 196, 199, 536, 541, 545, 548, 586, 589, 591, 596, 599. No passive residues for the DNA, but for MpARF2: 131, 134, 135, 140, 149, 187, 188, 531, 534, 535, 540, 549, 587, 588. After a HADDOCK run, four structures per cluster were used and aligned to the original MpARF2 structure using the *extra_fit* command with options *align* and *cycles = 0* in PyMOL. Since we assumed our MpARF2 mutations would cause minimal changes to the DNA structure, we selected the cluster with the least RMSD compared with the original (~0.5 RMSD averaged for all structures shown in this chapter). Representative HADDOCK structures were shown in the figures. All structural figures were made with PyMOL. Rotamers were shown using the *mutation* tool and hydrogen bonds using the *distance* command with *cutoff = 3.2* and *mode = 2*.

Supplementals

Supplemental table 1 | List of auxin-related genes in *M. polymorpha*.

ID	Gene Symbol
Mp1g07070	MpARF3
Mp1g09500	MpWIP
Mp1g12750	MpARF1
Mp2g02890	MpncARF
Mp2g24200	MpC2HDZ
Mp3g05910	MpR2R3-MYB21
Mp4g11820	MpARF2
Mp5g14320	MpTAA
Mp6g02750	MpTIR1
Mp6g05000	MpIAA
Mp8g07850	MpncIAA
Mp8g08780	MpYUC2

Supplemental table 2 | List of primers used in this study.

PrimerID	Sequence (5'-3')
Cloning primers	
JR375_MpARF1N_F	ATTCCCCGCCGTGCCGCA
JR376_MpARF1N_R	AGTATCACTTGCTGTCAATGTTTTGCAGAAAAAATCTGTAGGTTG
MpARF1N_mut*	ttgacagcaagtgatactAGTACTCATGGTGGTTTTTCAattccccgccgtgcc
JR377_MpARF1C_F	CTCACAACTGGATGGAGTG
JR378_MpARF1C_R	GTAGATGTGCCGGAAGTG
MpARF1C_mut*	cttccggcacatctacAGAGGTCAA
JR428_HiFi_F	gttacgcgtcgttccactgaGCGTCAGACCCCGTAGAAAAG
JR429_HiFi_R	TCAGTGGAACGACGCGTAAC
RT-qPCR	
MpACT7_F1	AGGCATCTGGTATCCACGAG
MpACT7_R1	ACATGGTCGTTCCTCCAGAC
MpAPT3_F1	CGAAAGCCCAAGAAGCTACC
MpAPT3_R1	GTACCCCCGGTTGCAATAAG
MpSAND_F1	GTTGATGTGTGGCACAAAGG
MpSAND_R1	CAGGCATACGGGAGAAAATG
JR171_MpARF1_F	CACAAAGATCAGGCTGGCAG
JR172_MpARF1_R	ACTTCCGACGGTGACAAGAT

* Shown are wild type sequences. Mutants were created by substituting the H146, R194, Q196, P197 and R199 codons.

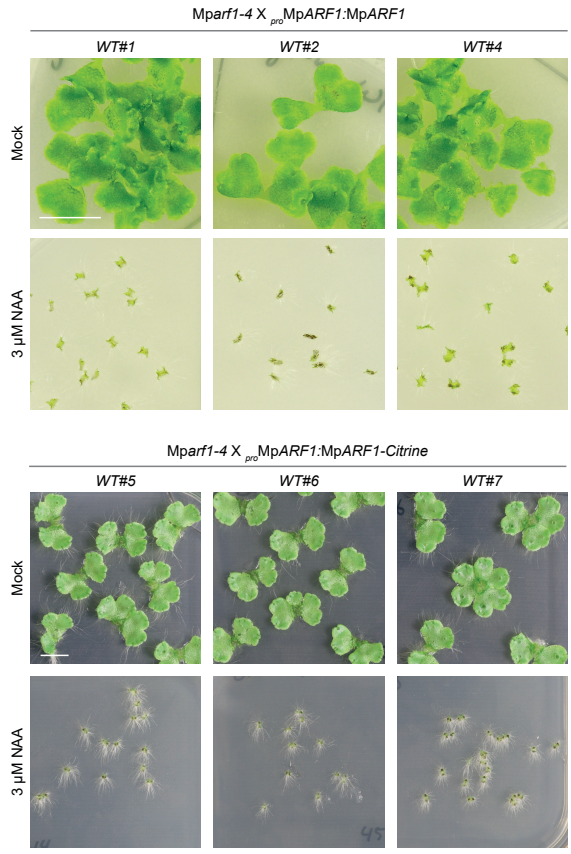


Fig. S1 | Growth phenotype of MpARF1-Citrine and non-Citrine lines. Fourteen day old gemmalings of Mp*arf1-4* transformed with either $_{pro}$Mp*ARF1:*Mp*ARF1* (upper panels) or $_{pro}$Mp*ARF1:*Mp*ARF1-Citrine* constructs grown without (mock) or with 3 µM NAA.

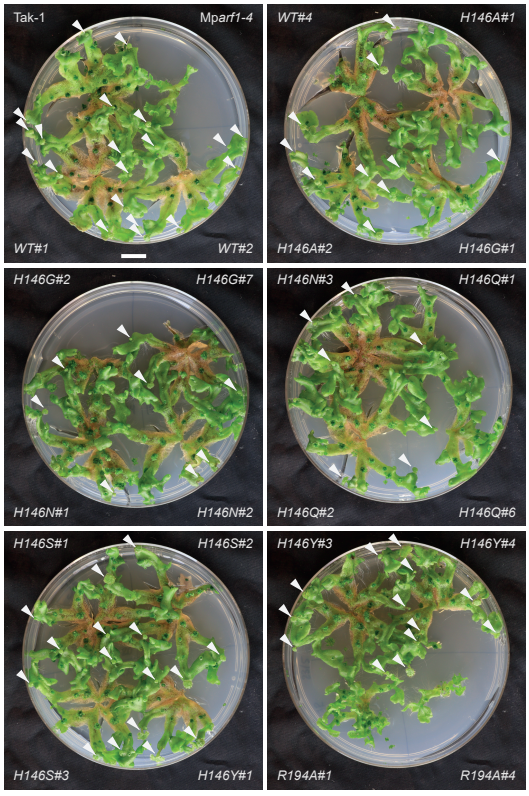


Fig. S2 | Restored antheridiophoregenesis in Mp*ARF1-H146* mutants. Topview of 7 week old thallus (one per transgenic line), with growing antheridiophores indicated with white arrows.

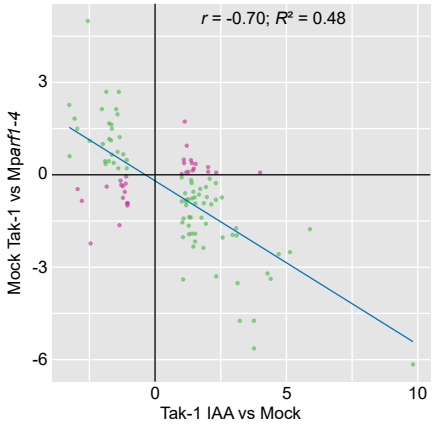


Fig. S3 | Correlation analysis IAA-mediated DEGs in Tak-1. Scatterplot and Pearson correlation (blue line) of 112 IAA-mediated genes in Tak-1, plotting the Log_2 FC of Tak-1 IAA versus Mock (x-axis) and the Tak-1 versus Mp*arf1-4* under mock (y-axis) conditions. Dots represent genes with either the same (magenta) or opposite (green) patterns.

Chapter

6

ARF-B3 domain DNA-binding conservation across ARF classes and streptophyte lineages

Juriaan Rienstra[1], Martijn de Roij[1*], Vanessa Polet Carrillo-Carrasco[1*], Jorge Hernández-García[1] and Dolf Weijers[1]

[1] Laboratory of Biochemistry, Wageningen University and Research, Wageningen, the Netherlands

[*]These authors contributed equally

Abstract

The AUXIN RESPONSE FACTOR (ARF) family of transcription factors can be divided into three classes – A, B and C – that are found in nearly all extant streptophytes. The ARF DNA-binding domain (DBD) contains a B3 domain that forms atomic contacts with TGTCNN auxin response elements (AuxREs) wherein N5 and N6 can be variable. Additionally, DBDs can homodimerize to bind two AuxRE sites with specific orientation and spacing. Orientation and spacing between AuxREs is an important component of ARF functional differentiation, as each class appears to have their own unique orientation and spacing preferences. A previous domain swap experiment with *Marchantia polymorpha*, a species with only one gene copy for each ARF class, showed that the DBDs of each class were functionally diverged *in vivo*. However, the A and B-ARFs of land plants universally exhibit the highest affinity to TGTCGG elements *in vitro*, which can also be bound by C-ARFs *in vitro*. It was therefore unclear whether the functional divergence of the *M. polymorpha* ARF DBDs was due to spacing or AuxRE binding preferences. Here, we explored functional and structural conservation and diversity of B3 domains. To that end, we complemented the *Marchantia polymorpha* null mutant Mp*arf1-4* with constructs of chimeric Mp*ARF1* wherein the B3 domain was exchanged to uncouple DBD spacing preferences from B3 AuxRE binding preferences. We found that the B3 domains of streptophyte algae A/B-ARFs, the *M. polymorpha* B-ARF and *Arabidopsis thaliana* A-ARFs could all functionally complement the endogenous Mp*ARF1* B3 domain, whereas the C-ARFs were unable to fully complement. This indicates that the similar AuxRE preferences of the A and B-ARFs were gained in the ancestral A/B-ARF before its duplication these preferences have been retained in extant streptophyte lineages, including in expanded A-ARF families from angiosperms. In contrast, the B3 domains of the C-ARFs have partially diverged from the A/B-ARFs and likely prefer other DNA sites.

Introduction

The AUXIN RESPONSE FACTOR (ARF) family of transcription factors (TFs) is widespread within the Streptophyte lineages (Carrillo-Carrasco *et al.*, 2023; Martin-Arevalillo *et al.*, 2019; Mutte *et al.*, 2018). The ARF family can be categorized into three classes that are found in nearly all extant lineages of land plants: A, B and C-ARFs. However, phylogenetic studies have revealed that ARFs originated in the common ancestor of streptophyte algae and land plants **(Fig. 1A)**. The inferred evolutionary scenario suggests that an ancestral proto-ARF duplicated, leading to the formation of a proto-A/B-ARF and a proto-C-ARF. Both classes are still present in extant streptophyte algae, albeit not with particular patterns of emergence or retention. A duplication of the proto-A/B-ARF gave rise to a last common ancestor of land plants with a single copy of A, B and C-ARFs, a state that is retained in the extant liverwort *Marchantia polymorpha* (Bowman *et al.*, 2017; Flores-Sandoval *et al.*, 2015; Kato *et al.*, 2015; Mutte *et al.*, 2018). These ARF families rapidly expanded in the angiosperms, with *Arabidopsis thaliana* having 5, 15 and 3 paralogs of A, B and C-ARFs respectively (Okushima *et al.*, 2005).

ARFs are modular proteins comprising a DNA-binding domain (DBD), a middle region (MR) that determines their transcriptional activity, and a Phox and Bem1 domain (PB1) that facilitates homo-, hetero- and oligomerization (Boer *et al.*, 2014; Korasick *et al.*, 2014; Nanao *et al.*, 2014; Ulmasov *et al.*, 1999a). The DBD is central to ARF function and can be structurally subdivided into subdomains (see Chapter 2): a dimerization domain (DD), a B3 domain that binds to auxin response elements (AuxRE), and a single alpha-helix, the B3-hinge, which connects the DD and B3 domains (Chapter 2, Chapter 3). The DD enables the DBD to form head-to-head homodimers that can bind two AuxRE elements in an inverted repeat (IR) orientation, a dimerization interface conserved between B-ARFs from *A. thaliana* (Arabidopsis) and *M. polymorpha* (Boer *et al.*, 2014; Kato *et al.*, 2020). However, since A-ARFs can additionally bind direct repeat (DR) and everted repeat (ER) motifs, other interaction interfaces likely exist (Chapter 4) (Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). For each orientation, ARFs exhibit specific spacing preferences between two AuxRE. Whereas angiosperm A-ARFs can bind IR motifs spaced by 5 to 9 base pairs, B-ARFs are limited to IR motifs with 7 or 8 base pair spacing (Boer *et al.*, 2014). In contrast, C-ARFs can bind IR motifs with 4 to 9 base pairs spacing, a preference conserved between C-ARFs from the unicellular streptophyte alga *Chlorokybus atmophyticus* and Arabidopsis (Martin-Arevalillo *et al.*, 2019). Thus, both the orientation and spacing between two AuxRE are significant factors contributing to the functional divergence between ARF classes.

The ARF B3 domain facilitates DNA-binding by forming base-specific atomic interactions with AuxREs. This AuxRE consists of the nucleotides TGTCNN, with the core sequence TGTC, and the variable nucleotides N5 and N6 (Chapter 2, Chapter 3) (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Ulmasov *et al.*, 1999b; Ulmasov *et al.*, 1995). Land

plant A and B-ARFs from *M. polymorpha* (MpARF1), Arabidopsis (AtARFs) and maize (*Zea mays*; ZmARFs) show the highest affinities to TGTC elements ending with GG (TGTCGG), although they may bind other TGTCNN elements with lower affinities *in vitro* and *in vivo* (Chapter 4) (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014; Freire-Rios *et al.*, 2020; Galli *et al.*, 2018; Lieberman-Lazarovich *et al.*, 2019; Zemlyanskaya *et al.*, 2016). C-ARFs have also been shown to bind TGTCGG and TGTCTC elements *in vitro* (Martin-Arevalillo *et al.*, 2019), but nothing is known about their preferences. Structural studies of the B-ARF AtARF1 revealed that four strongly conserved residues – R181, Q183, P184 and R186 – are crucial for binding to the core-TGTC, as alanine mutations of these residues strongly impair the ability of A-ARFs AtARF5 and MpARF1 to complement their respective null mutants (Chapter 5) (Boer *et al.*, 2014). Binding to the variable N5 and N6 nucleotides is achieved by a single reside (H136) in AtARF1, with its conformational flexibility determining which elements are bound with high affinity (TGTCGG) or lower affinity (e.g., TGTCTC) (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020). These DNA-binding residues are strongly conserved between ARF classes, between species, and across all lineages of Streptophytes (Chapter 3) (Boer *et al.*, 2014; Mutte *et al.*, 2018). Nevertheless, variation for B3 domain residues outside of the DNA-binding residues has been reported previously (Boer *et al.*, 2014), variation that has been shown to alter DNA-binding in other transcription factors (Aditham *et al.*, 2021; Dorrity *et al.*, 2018; Koulechova *et al.*, 2015). Additionally, other transcription factor families can share similar high-affinity sites but exhibit variation in their preferences for low and medium affinity sites (Badis *et al.*, 2009; Berger *et al.*, 2008; Crocker *et al.*, 2015). This raises the possibility that in species with expanded ARF families like the angiosperms, all their ARFs may strongly prefer TGTCGG elements (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014; Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019), but it remains unknown if variation outside of their DNA-binding residues affects their preferences for low to medium affinity elements.

ARF DNA-binding therefore comprises two components: binding to AuxRE elements and spacing between two AuxREs. A recent study showed that the DBDs of the B and C-ARFs from *M. polymorpha* were functionally different from the A-ARF, which was attributed to different DNA-binding preferences (Kato *et al.*, 2020). However, whether this difference was due to spacing preferences or alternative preferences to TGTCNN elements is unknown. To effectively test whether ARF B3 domains are capable of binding similar elements, it is necessary to uncouple both components.

In this chapter, we investigated whether B3 domains from ARFs of different Streptophyte lineages and classes are functionally similar. To assess this, we performed complementation assays using the *M. polymorpha* Mp*arf1-4* null mutant in combination with chimeric MpARF1 constructs wherein the B3 domain was swapped with B3 domains from other ARFs. This approach allowed us to evaluate the functional similarity of B3 domains and thereby infer whether DNA-binding to TGTCNN elements was comparable.

We assessed B3 domains from B-ARFs and C-ARFs (Mp*ARF2* and Mp*ARF3*, respectively), the expanded A-ARF family of Arabidopsis, and the A/B-ARF and C-ARF from extant charophyte algae. Overall, our results indicate that the B3 domain of most ARFs can functionally complement, suggesting that the DNA-binding to TGTCNN elements is conserved across multiple Streptophyte lineages and ARF classes.

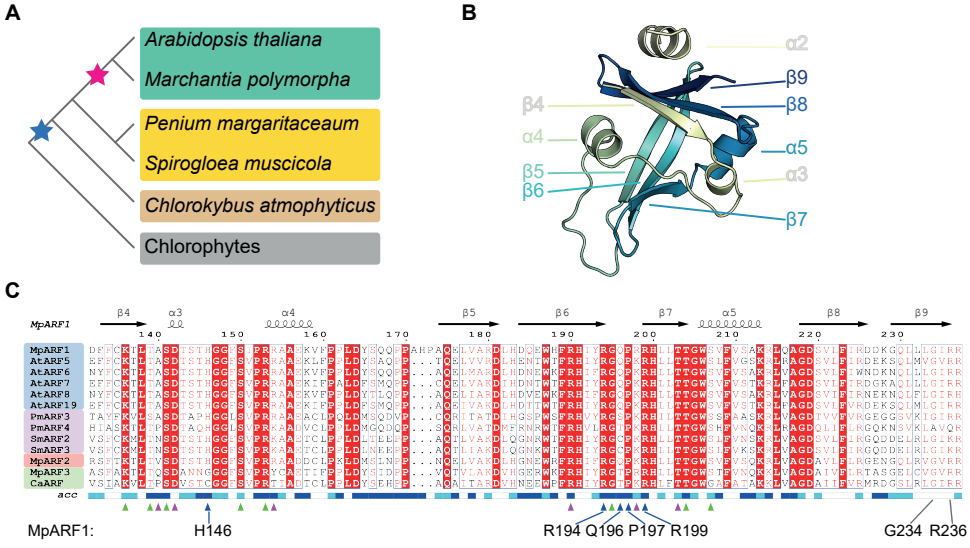


Fig. 1 | Alignment of ARF B3 domains used in this study. (A) Schematic phylogeny of species used in this study and inferred ARF evolutionary scenario. The colored block represents Streptophyte lineages, from top to bottom: land plants (turquoise); Zygnematophyceae (gold); Chlorokybophyceae (brown); and Chlorophytes as outgroup (grey). Star represent the first occurrence of proto-A/B and proto-C ARFs (blue; see Carrillo-Carrasco *et al*., 2023), and the first occurrence of A and B-ARFs (magenta). **(B)** Predicted structure of the *M. polymorpha* MpARF1 B3 domain and α-helix2 (homologous to AtARF1 α-helix1; Chapter 3). Structure was predicted using AlphaFold2 (Jumper *et al*., 2021; Varadi *et al*., 2021). **(C)** Alignment of A (blue), B (red), A/B (purple) and C (green) ARFs. Residue numbering is based on MpARF1. Helices and arrows on top indicate α-helices and β-sheets respectively in the predicted structure of MpARF1 from panel B. Row acc indicates accessibility of each residue in the predicted MpARF1 structure: accessible (dark blue), intermediate accessible (cyan) and buried (white). Triangles indicate residues involved in DNA-binding on the: major groove (blue), phosphate backbone (green) or via water-mediated hydrogen bonds (purple), as in Chapter 3.

Results

Variation in B3 domains

In Chapter 3, we extensively compared the conservation of residues in the DBD and B3 domains between different lineages of the streptophytes, including four algal species. We concluded that the overall B3 structure was conserved across all lineages. Here, we took a subset of these ARFs to test their B3 domains in complementation assays **(Fig. 1B, 1C)**. We used the proto-A/B-ARFs from *Penium margaritaceaum* (*PmARF3* and *PmARF4*) and *Spirogloea muscicola* (*SmARF2* and *SmARF3*), and the proto-C-ARF from *Chlorokybus atmophyticus* (*CaARF*). With these ARFs, we cover the extant grades of streptophyte algae **(Fig. 1A)**. Furthermore, we used the B3 domains of Mp*ARF2* and Mp*ARF3* to represent land plant B and C-ARFs respectively. Finally, we used B3 domains of all A-ARFs from Arabidopsis to represent an extended A-ARF family from angiosperms.

To understand the variability between the B3 domains of this subset of ARFs we produced a multiple sequence alignment **(Fig. 1B and 1C)**. Out of the five residues that enter the major groove, three of the functionally critical residues – MpARF1-R194, P197 and R199 – are conserved, and the two residues that tolerate other residues without affecting function – MpARF1-H146 and Q196 (Chapter 5) – are more variable. For H146, all A, B and A/B-ARFs used in this study have a histidine, suggesting these ARFs likely share the same preference for TGTCGG elements. With the C-ARFs in comparison, while MpARF3 has a C-ARF-specific, conserved glycine (G) at this position (Chapter 3, Chapter 5), CaARF has a cysteine (C) at this position. These variants could affect preferences for certain TGTCNN elements, since an H146G substitution in Mp*ARF1* caused reduced gemma cup production in *M. polymorpha* (Chapter 5). Q196 was likewise conserved in the A and B-ARFs of Arabidopsis and *M. polymorpha* **(Fig. 1C)**, while both C-ARFs have a C-ARF-specific, conserved threonine (T; Chapter 5). The algal ARFs show differences however, with *PmARF3* and *PmARF4* having a serine (S) and threonine at this position respectively, and *SmARF3* and *SmARF4* both having a cysteine **(Fig. 1C)**. Since Q196 forms base-specific hydrogen bonds via its peptide backbone and substitutions on this residue do not affect Mp*ARF1* function (Chapter 5), it is unclear if and how these variants might affect DNA-binding. Finally, the other residues that are not involved in DNA-binding showed much higher variability among the ARFs. Therefore, based on the conservation of the main three functionally critical DNA-binding residues, it was expected that these B3 domains could bind TGTCNN elements.

Functional complementation by B3 domains of B and C-ARFs

Previously, Kato *et al.* (2020) used a complementation assay of the *M. polymorpha* Mp*arf1-4* null mutant and showed the DBDs of Mp*ARF2* and Mp*ARF3* (i.e., respective DBD fused to MR and PB1 domain of Mp*ARF1*) were functionally different to the DBD

of Mp*ARF1*. Here, to answer if DNA-binding *per se* was different between the ARF classes, we uncoupled spacing preferences and AuxRE binding by using B3 swaps. For

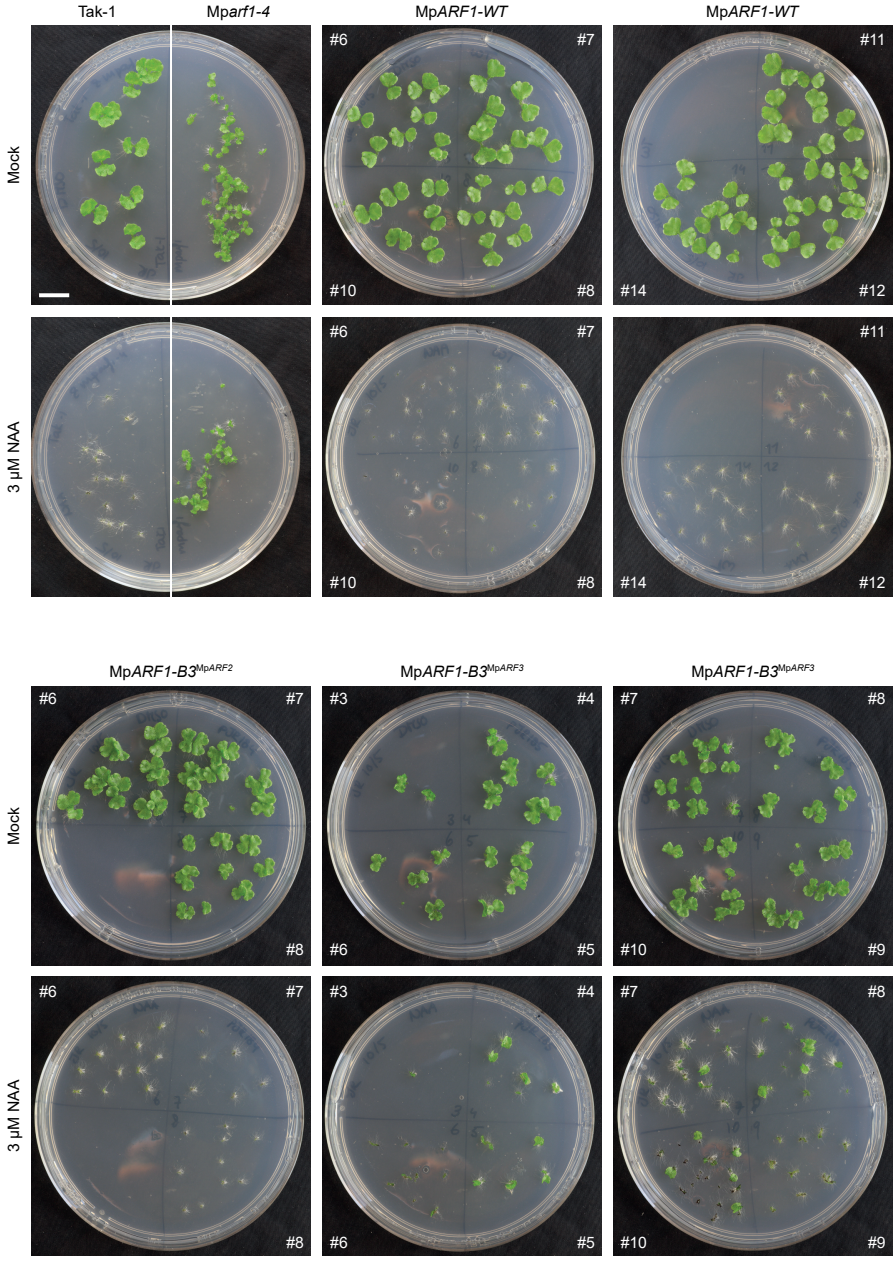


Fig. 2 | Growth phenotypes of B3 chimeras with the MpARF2 and MpARF3 B3 domains. Fourteen-day-old gemmalings of Tak-1, Mp*arf1-4* and Mp*arf1-4* transformed with $_{pro}$Mp*ARF1:*Mp*ARF1* constructs grown on medium without (mock) and with 3 µM NAA. Construct without mutations is termed Mp*ARF1-WT* and with B3 swaps as indicated. Numbers inside each panel correspond to individual transgenic lines. The scale bar is 10 mm.

this, we transformed Mp*arf1-4* with constructs of either wild-type Mp*ARF1* fused with a c-terminal *Citrine* ($_{pro}$Mp*ARF1:*Mp*ARF1-WT-Citrine*; hereafter Mp*ARF1-WT*) or constructs wherein the B3 domain was exchanged (Mp*ARF1-B3*$^{Mp ARF2}$ or Mp*ARF1-B3*$^{Mp ARF3}$). As previously observed (Ishizaki *et al.*, 2012; Kato *et al.*, 2015; Kato *et al.*, 2017; Sugano *et al.*, 2014), gemmalings of Takaragaike-1 (Tak-1) exhibit severe growth impairment and excessive rhizoid proliferation when grown on medium containing the synthetic auxin 1-Naphthaleneacetic acid (NAA), whereas gemmalings of the Mp*arf1-4* mutant do not respond to NAA but displayed equal growth impairment under mock and NAA conditions **(Fig. 2)**. Gemmalings of Mp*arf1-4* transformed with Mp*ARF1-WT* constructs are comparable in size to Tak-1 under mock conditions and exhibit rhizoid overproliferation under NAA conditions, indicating the Mp*ARF1-WT* construct fully complements **(Fig. 2)** (Kato *et al.*, 2017). For the Mp*ARF2* swap, we found the Mp*ARF1-B3*$^{Mp ARF2}$ construct to fully complement Mp*arf1-4* in all seven independent lines **(Fig. 2)**. The Mp*ARF2* DBD-swap was previously reported to only partially rescue (Kato *et al.*, 2020). These results therefore indicate that whereas the Mp*ARF1* and Mp*ARF2* DBDs are functionally close but unequal, their B3 domains are functionally comparable. For the Mp*ARF3* swap, gemmalings of Mp*arf1-4* transformed with the Mp*ARF1-B3*$^{Mp ARF3}$ construct were overall larger than Mp*arf1-4* under mock conditions. Under NAA conditions however, while the majority of gemmalings responded to NAA by inhibition of growth and rhizoid overproliferation, a minority showed an intermediate response, i.e. being slightly inhibited in growth without rhizoid overproliferation **(Fig. 2)**. These results indicated the Mp*ARF3* B3 domain function is only partially similar to Mp*ARF1*, thereby indicating functional divergence of the respective B3 domains. Since the Mp*ARF3* DBD was reported to not complement Mp*arf1-4* (Kato *et al.*, 2020), we conclude that A and C-ARFs functionally diverged at both the B3 and DBD levels. To test whether misexpression or mislocalization of the chimeric protein was the cause for the partial complementation, we investigated the protein localization in gemmae using the C-terminally fused Citrine **(Fig. 3)**. We found that distribution of the Citrine signal across the gemmae and the nuclear localization were comparable between Mp*ARF1-WT*, Mp*ARF1-B3*$^{Mp ARF2}$, and Mp*ARF1-B3*$^{Mp ARF3}$. We could therefore attribute the partial complementation of Mp*ARF1-B3*$^{Mp ARF3}$ to intrinsic properties of the protein, rather than misexpression or mislocalization. Combined, taking the inferred evolutionary trajectory of the A, B and C-ARFs into account, these results indicate that: i) the B3 domains of the A and B-ARFs remained functionally similar after the duplication of an A/B-like ancestor, and ii) that the B3 domains of A/B and C-ARFs functionally diverged.

Functional complementation by Arabidopsis A-ARFs orthologs of Mp*ARF1*

In Arabidopsis, an ancestral A-ARF gave rise to five paralogues (AtARF5, 6, 7, 8 and 19) with redundant, overlapping and unique roles in plant growth and development,

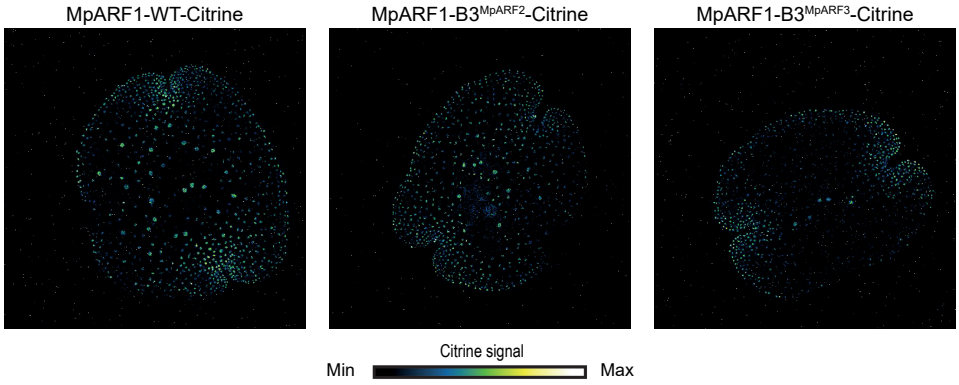


Fig. 3 | MpARF1-Citrine fusion protein localization and expression in gemmae. Gemmae of Mp*arf1-4* transformed with $_{pro}$Mp*ARF1:*Mp*ARF1-Citrine* (Mp*ARF1-WT*) or with swapped B3 domains of Mp*ARF2* or Mp*ARF3* (Mp*ARF1-B3*^{MpARF2} or Mp*ARF1-B3*^{MpARF3} respectively). Each image is representative of at least three gemmae.

indicating functional divergence (Hardtke *et al.*, 2004; Kato *et al.*, 2020; Okushima *et al.*, 2007; Okushima *et al.*, 2005; Rademacher *et al.*, 2012; Rademacher *et al.*, 2011). Whether this is due to differences in the B3 domains is unknown. Since these ARFs are orthologs of Mp*ARF1*, we examined whether the B3 domains of the Arabidopsis A-ARFs are functionally similar to Mp*ARF1*. Gemmalings from Mp*arf1-4* transformed with any of the Arabidopsis A-ARF swaps – Mp*ARF1-B3*^{AtARF5}, Mp*ARF1-B3*^{AtARF6}, Mp*ARF1-B3*^{AtARF7}, Mp*ARF1-B3*^{AtARF8}, and Mp*ARF1-B3*^{AtARF19} – were fully complemented **(Fig. 4)**. The size of the gemmalings of the B3 chimeras were comparable in size to Tak-1 and Mp*ARF1-WT* under mock conditions and showed a uniform response to NAA. Therefore, the Arabidopsis A-ARF B3-domains are functionally similar to Mp*ARF1*, indicating the conservation of an ancestral function between these B3 domains of land plant A-ARFs.

Functional complementation by B3 domains of streptophyte algae

Both A/B-ARFs and C-ARFs have been found in extant charophyte algae (Martin-Arevalillo *et al.*, 2019; Mutte *et al.*, 2018; One Thousand Plant Transcriptomes, 2019), of which the C-ARF *CaARF* was shown to bind TGTCTC and TGTCGG elements *in vitro* (Martin-Arevalillo *et al.*, 2019). We performed complementation assays to test if these algal B3 domains were functionally similar to land plant ARFs **(Fig. 5)**. From the algal A/B-ARFs we noted the Mp*ARF1-B3*^{PmARF3} and Mp*ARF1-B3*^{SmARF3} constructs fully complemented Mp*arf1-4* under mock and NAA conditions, except that one of the four Mp*ARF1-B3*^{PmARF3} lines failed to complement, possibly due to transgene silencing or low expression. Based on the consistency of the complementation for the other lines, we concluded that the B3 domains of *PmARF3* and *SmARF3* were functionally similar to that of Mp*ARF1*. For the other two constructs, Mp*ARF1-B3*^{PmARF4} and Mp*ARF1-B3*^{SmARF2}, Mp*ARF1-B3*^{PmARF4} failed to complement under both mock and NAA conditions, whereas Mp*ARF1-B3*^{SmARF2}

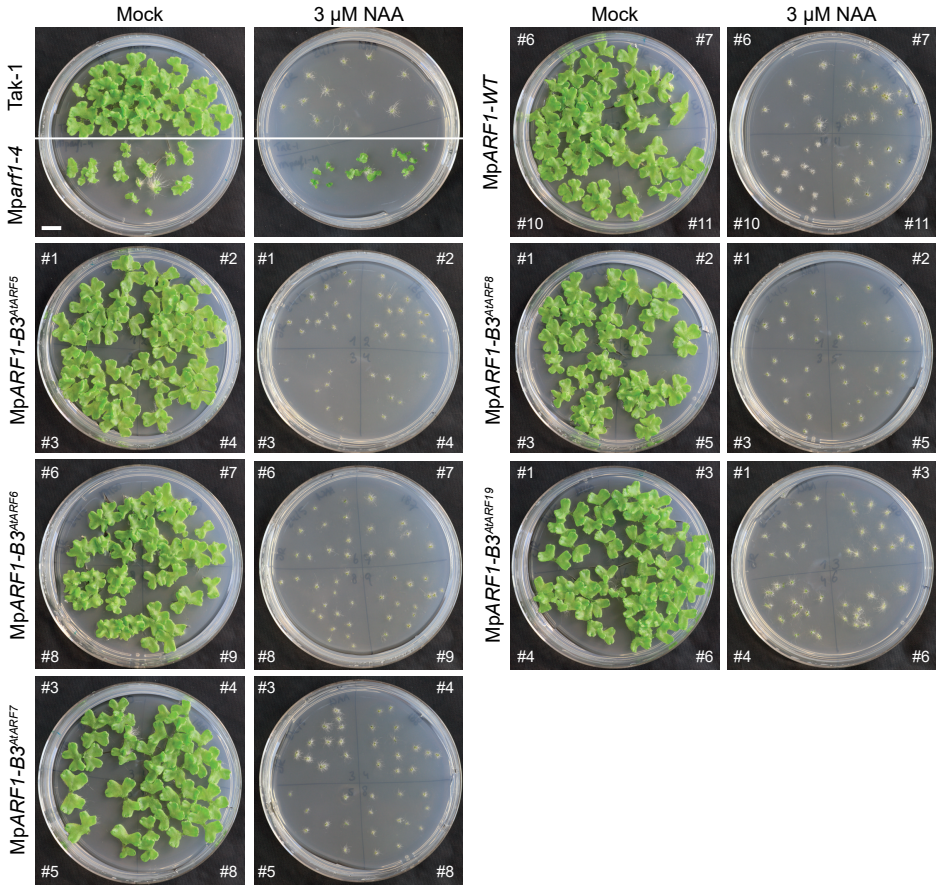


Fig. 4 | Growth phenotypes of B3 chimeras with Arabidopsis A-ARFs. Fourteen-day-old gemmalings of Tak-1, Mp*arf1-4* and Mp*arf1-4* transformed with $_{pro}$Mp*ARF1:*Mp*ARF1* constructs grown on medium without (mock) and with 3 µM NAA. Construct without mutations is termed Mp*ARF1-WT* and with B3 swaps as indicated. Numbers inside each panel correspond to individual transgenic lines. Note that pictures of this figure and Fig. 5 were performed in the same experiment. The scale bar is 10 mm.

partially complemented, with gemmalings being slightly larger than Mp*arf1-4* under mock conditions and showing variable responses to NAA. It therefore shows that some algal A/B-ARFs retain a conserved function similar to the B3 domain of Mp*ARF1*, whereas others might have functionally diverged or acquired mutations that distort the original function. Finally, transforming Mp*arf1-4* with a B3-chimera of the C-ARF *CaARF* (Mp*ARF1-B3*^{CaARF}) resulted in partial complementation **(Fig. 5)**. All four lines were comparable in size to Mp*arf1-4* under mock conditions. Under NAA conditions, two lines showed a response to NAA, indicating restoration of auxin signaling, while two lines were unresponsive. Given that independent lines transformed with Mp*ARF1-WT* were uniformly responding to NAA **(Fig. 2, 4, 5)**, it shows the B3 of *CaARF* is only partially functionally comparable

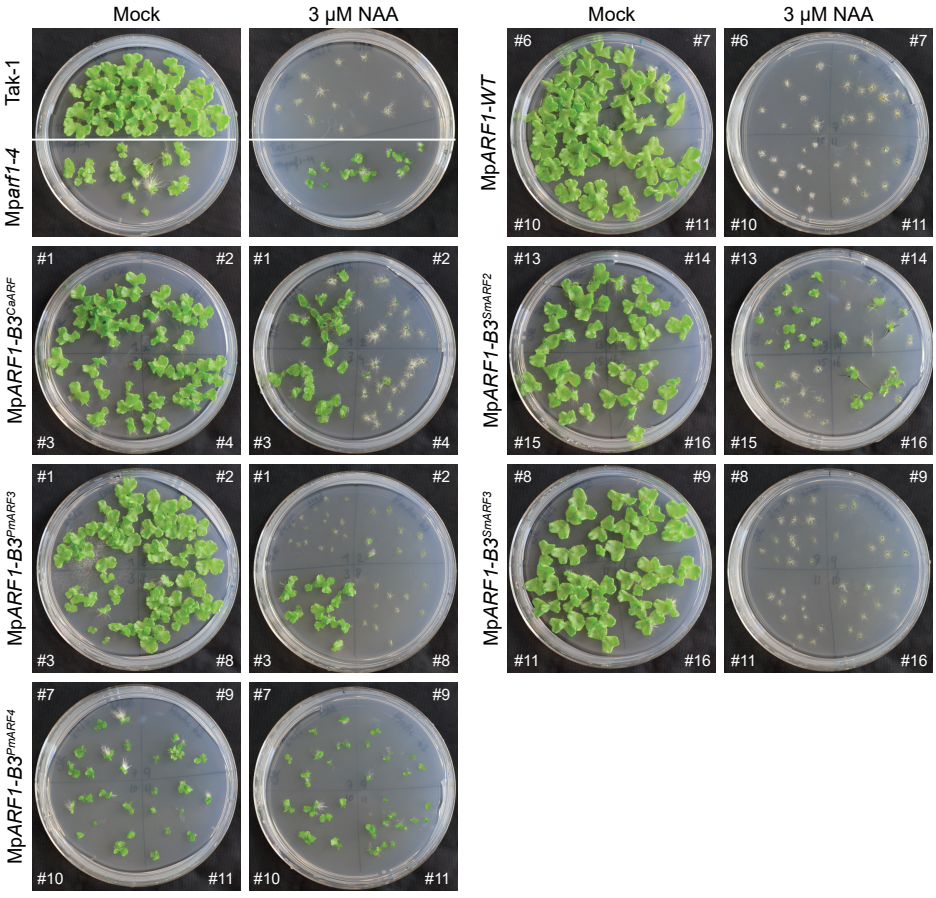


Fig. 5 | Growth phenotypes of B3 chimeras with algal ARFs. Fourteen-day-old gemmalings of Tak-1, Mp*arf1-4* and Mp*arf1-4* transformed with $_{pro}$Mp*ARF1:*Mp*ARF1* constructs grown on medium without (mock) and with 3 µM NAA. Construct without mutations is termed Mp*ARF1-WT* and with B3 swaps as indicated. Numbers inside each panel correspond to individual transgenic lines. Note that pictures of this figure and Fig. 4 were performed in the same experiment. The scale bar is 10 mm.

to Mp*ARF1*. Together, these results indicate that the B3 domains of some A/B-ARFs from streptophyte algae can fulfill similar functional roles as the B3 domain of Mp*ARF1*, while the C-ARFs functionally diverged.

Structural view on tolerated residue variation in the B3 domains

All B3 chimeras except Mp*ARF1-B3*^{PmARF4} were able to complement Mp*arf1-4* at least partially, despite the presence of extensive variation on the residue level between the B3 domains of Mp*ARF1* and the other ARFs **(Fig. 1B)**. To understand where these variations occurred within the B3 structure, we tabulated all positions at which other residues occurred between Mp*ARF1* and all other B3 domains except *PmARF4* **(Fig. 6A)**.

Surprisingly, the tolerated variation occurs in all sub-regions of the Mp*ARF1* B3 domain. For instance, variation in the hydrophobic core could alter conformational stability and consequently DNA-binding (Koulechova *et al.*, 2015; Lee *et al.*, 2019; Ventura *et al.*, 2002), yet these residues exhibit variability between the B3 domains that do not appear to impede their function *in vivo*. Furthermore, many residues in the B3-hinge region were variable across the B3 domains. **(Fig. 6A)**. These regions are therefore relatively tolerable to variation.

The B3 domain of *PmARF4* was unable to complement Mp*arf1-4*, though the conservation of the DNA-binding residues **(Fig. 1B)** would suggest it has similar DNA-binding as MpARF1. We reasoned that some crucial residues must be altered only in *PmARF4* and not in the other B3 domains. We therefore looked at each individual residue that differed between Mp*ARF1* and *PmARF4*. If that same residue was changed to a similar physiochemical residue in a partial complementing B3, or if the change in *PmARF4* also occurred in a partially complementing B3, we tabulated it as a likely acceptable change. The remaining different residues between Mp*ARF1* and *PmARF4* were deemed detrimental **(Fig. 6B)**. That process left two residues in the B3-hinge region that were solely different in *PmARF4*: MpARF1-G234 and R236 **(Fig. 6B)**. All the other B3 domains share these residues, apart from *PmARF4*, which has an alanine (A) and glutamine (Q) at these positions respectively. Closeup in the AlphaFold predicted structure of MpARF1 (Chapter 3) (Jumper *et al.*, 2021; Varadi *et al.*, 2021) revealed that G234 is buried by hydrophobic residues and shares the pocket with residue C35, which stems from α-helix 2 **(Fig. 1B, 6C)**. This glycine and cysteine are completely conserved in ARFs from all three classes in the phylogenetic tree of Mutte *et al.* (2018; not shown) and G234 in particular is amongst the highest conserved B3-hinge residues in a larger sample of ARFs across the plant kingdom (**Fig. 6D**; Consurf score 9; Ashkenazy *et al.*, 2016). The role of C35 (Consurf score 8) in the ARFs is unclear, as depending on their microenvironment cysteines can be classified as either highly hydrophobic buried residues or highly polar and reactive solvent exposed residues (Marino and Gladyshev, 2010; Mazmanian *et al.*, 2016). G234 might create enough space to fit C35 in this pocket in MpARF1, and the alanine from PmARF4 could disrupt this space and lead to misfolding of the full DBD. R236 forms electrostatic interactions with E313 in the predicted MpARF1 structure and might fulfil a structural role **(Fig. 6E)**. The neutral, polar glutamine (Q) of *PmARF4* might create weaker interactions than the conserved R236 as a glutamine could only form hydrogen bonds and not electrostatic interactions. Thus, the likely explanation of failure of the *PmARF4* B3 domain to complement is due to two variants for key residues of the B3-hinge region. Based on these results we conclude that the B3 domains, if they retain the conserved DNA-binding residues, is relatively tolerable to mutations occurring in the B3-hinge or hydrophobic core that may affect conformational stability but do not negatively impact folding. Only specific mutations that disrupt or weaken the B3 structure or B3-hinge, such as G234A and R236Q, can impair function.

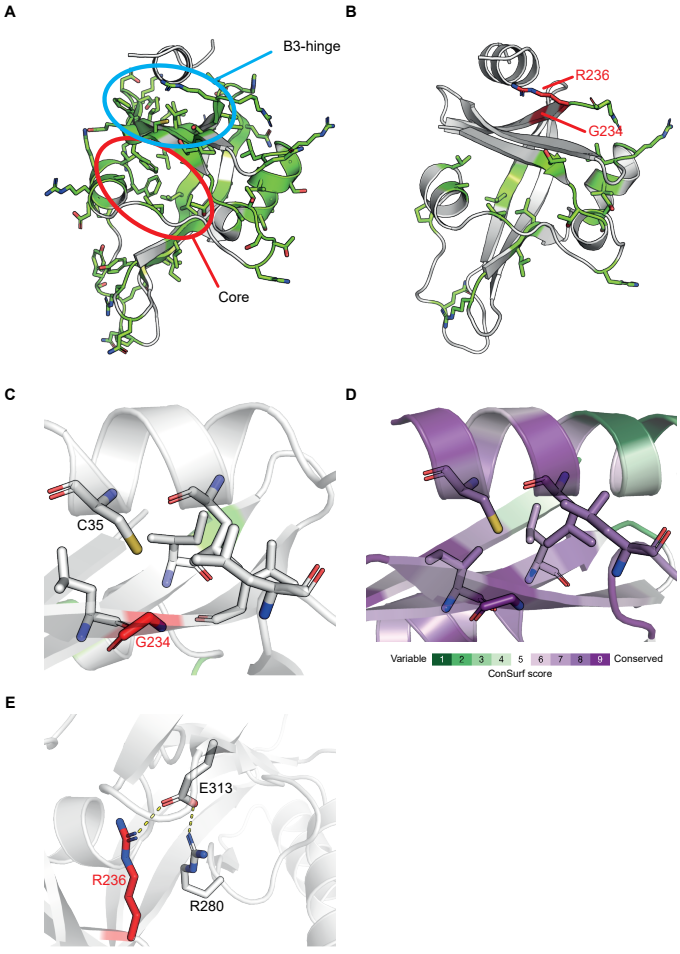


Fig. 6 | Tolerated variation between B3 domains. (A) Predicted structure of the MpARF1 B3 domain and α-helix2 (homologous to AtARF1 α-helix1, Chapter 3) as a white cartoon. Green cartoon and sticks: positions on which variation was found in the other B3 domains that partially or fully complemented. Circles highlight residues from the hydrophobic core (red) and B3-hinge (blue) respectively. **(B)** Predicted structure of the MpARF1 B3 domain and α-helix2 as white cartoon, with variable positions in *PmARF4* as green or red sticks and cartoon. Green are residues that are different in *PmARF4* but which were found in the other B3 domains as well in A, whereas red are residues that are only different in *PmARF4*. **(C)** Closeup of B, showing the hydrophobic residues surrounding MpARF1-C35 and G234 (red stick). **(D)** Closeup of the same regions as C, but in the AtARF1 ConSurf analysis (PDB: 6ycq; see Chapter 3 for details; (Ashkenazy *et al.*, 2016)). ConSurf conservation scores per residue range from 1-9 (green to purple), higher score is more conserved. **(E)** Closeup of B, showing the electrostatic interactions (yellow dashes) between R236 or R280 with E313 in MpARF1. In all panels, the coloring of stick residues follows standard practices: blue, nitrogen; red, oxygen; yellow, sulfur; carbons are colored for specific purposes. Predicted structure of MpARF1 was made with AlphFold2 (Jumper *et al.*, 2021; Varadi *et al.*, 2021).

Discussion

Based on *in vitro* DNA-binding assays and the conservation of DNA-binding residues, it was hypothesized that ARFs from all classes and across species would bind TGTCGG elements with high affinity (Chapter 3) (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014; Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). However, limited information was available on affinities towards other TGTCNN elements, raising the possibility that ARFs subspecialized on medium affinity sites, as was observed for Hox and Homeodomain transcription factor families (Berger *et al.*, 2008; Crocker *et al.*, 2015; Noyes *et al.*, 2008a). DBD-swaps in *M. polymorpha* suggested differences in DNA-binding preferences between A, B and C-ARFs. However, due to unique bipartite spacing preferences across classes (Boer *et al.*, 2014; Galli *et al.*, 2018; Stigliani *et al.*, 2019), it remained unclear whether these differences were attributable to spacing or distinct TGTCNN preferences. In this chapter, we aimed to decouple spacing preferences from DNA-binding to investigate functional similarities among ARF B3 domains.

Here, we found that the B3 domains of A-ARFs from Arabidopsis, the B-ARF Mp*ARF2* and the A/B-ARFs from *P. margaritaceaum* and *S. muscicola* could almost all (except *PmARF4*) partially or fully functionally complement Mp*ARF1*. First, we found the B3 domain of Mp*ARF2* was able to fully functionally complement Mp*ARF1*, suggesting full overlap of TGTCNN preferences between the A and B-ARFs. In contrast, the full Mp*ARF2* DBD was only partially complementing the DBD of Mp*ARF1* (Kato *et al.*, 2020), implying that intrinsic properties of the Mp*ARF2* DD and not the B3 are different from Mp*ARF1*. In angiosperms, B-ARFs have different spacing preferences and homodimerization affinities compared to A-ARFs (Boer *et al.*, 2014; Fontana *et al.*, 2023; Freire-Rios *et al.*, 2020; Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). It will therefore be essential to elucidate the DNA-binding preferences of MpARF2 to find if it competes with MpARF1 on specific sites, as is suggested (Kato *et al.*, 2020), and if it can bind MpARF2-unique sites. Second, the Arabidopsis A-ARFs tested each have unique, overlapping and redundant functions *in vivo* (Hardtke *et al.*, 2004; Okushima *et al.*, 2007; Okushima *et al.*, 2005; Rademacher *et al.*, 2012; Rademacher *et al.*, 2011). Notably, none of these studies used DBD swaps, instead using promoter swaps and mutant lines, and were therefore unable to pinpoint differences in TGTCNN preferences. Here, we found that all B3 domains of Arabidopsis A-ARFs can functionally complement MpARF1, indicating that all these domains retain similar TGTCNN binding preferences that must have been shared with the last common ancestor of land plants. Nevertheless, despite the Arabidopsis A-ARFs retaining an ancestral DNA-binding preference, we cannot exclude the possibility that these ARFs additionally subspecialized for their DNA-binding in Arabidopsis, as their cell-specific expression profiles might suggest (Rademacher *et al.*, 2011). Additionally, in Arabidopsis TGTCNN elements have been associated with particular processes in Arabidopsis (e.g., TGTCTT with phosphate metabolism (Lieberman-Lazarovich *et al.*,

2019; Zemlyanskaya *et al.*, 2016)), which could suggest that some cell-type specific ARFs specialized on these elements. Whereas in *M. polymorpha*, Mp*ARF1* would be solely responsible for all these processes and therefore may use a smaller subset of TGTCNN elements. To further progress our understanding of any subspecialization of DNA-binding in the Arabidopsis ARFs, one could use the modular nature of the DBDs as capitalized on here and in other studies (Kato *et al.*, 2020) and perform DBD/B3 swaps between the Arabidopsis A-ARFs to test the extend of conservation/specialization. Finally, the algal A/B ARFs have been speculated to bind similar DNA-elements as land plant ARFs based on homology models and conservation of DNA-binding residues (Chapter 3) (Kato *et al.*, 2020; Martin-Pizarro *et al.*, 2019; Mutte *et al.*, 2018). Since two of the algal A/B-ARF B3 domains (*PmARF3* and *SmARF2*) were able to functionally complement MpARF1, this indicates that these algal B3 domains can indeed bind similar DNA-elements. To conclude, these results indicate that similar TGTCNN binding preferences have been maintained throughout all extant streptophyte A-ARF lineages, and likely all B-ARF lineages, and suggests that these preferences were a feature of the last common ancestor of the A/B ARFs.

Limited knowledge exists regarding the DNA-binding preferences of C-ARFs, and our results suggest that DNA-binding preferences are likely different from A-ARFs. A single study indicates that AtARF10 and CaARF can bind both TGTCTC and TGTCGG elements in IR motifs with spacing of 4 to 9 base pairs as homodimers *in vitro* (Martin-Arevalillo *et al.*, 2019), which are similar elements and motifs A-ARFs can bind (Boer *et al.*, 2014; Galli *et al.*, 2018; Martin-Arevalillo *et al.*, 2019; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). Nevertheless, in *M. polymorpha* the DBD of Mp*ARF3* was found to be functionally distinct from Mp*ARF1* (Kato *et al.*, 2020). Because the histidine residue (MpARF1-H146) responsible for specifying binding preferences to N5 and N6 of the TGTCNN elements is replaced with a conserved glycine in C-ARFs (Chapter 3) (Martin-Arevalillo *et al.*, 2019; Mutte *et al.*, 2018), C-ARFs may display unique preferences for TGTCNN elements. First, we demonstrated here that B3 domains of Mp*ARF3* or *CaARF* only partially complemented Mp*ARF1*, which implies functional divergence and in turn suggests there are unique TGTCNN preferences between A-ARFs and C-ARFs. This could be further investigated by performing DNA-binding assays with C-ARFs, preferably methods that could measure affinities to lower and medium sites as well, such as PBM or SELEX (Godoy *et al.*, 2011; Jolma *et al.*, 2010; Smaczniak *et al.*, 2017). Second, although it was previously shown that, after the duplication event of the proto-C-ARF, the A/B-ARFs and C-ARFs acquired distinct spacing preferences (Martin-Arevalillo *et al.*, 2019), we show that distinct TGTCNN binding preferences (as inferred from the partial complementation) are another aspect of their functional divergence. Still, by performing B3 swaps with Mp*ARF1*, this is only viewed from the position of the A-ARFs and cannot exclude the possibility that A-ARFs can completely functionally replace C-ARFs. It will therefore be interesting to complement Mp*arf3* mutants with Mp*ARF3* with DBD and

B3 swaps with A-ARFs and B-ARFs. Furthermore, our assay could not indicate whether the B3 domains of *CaARF* and Mp*ARF3* are functionally similar to each other, which the Mp*ARF3-B3* swaps could reveal. Finally, the results shown here corroborate and extend previous studies and demonstrate that C-ARFs functionally diverged from A-ARFs on both spacing and AuxRE preferences.

In Chapter 3, we found that half of the B3-hinge residues were strongly conserved across the plant kingdom, while the other half exhibited high variability, potentially indicating differences in interdomain flexibility between ARFs. This interdomain flexibility was previously speculated to be causal for differences in bipartite spacing preferences between AtARF1 and AtARF5 (Boer *et al.*, 2014). However, although we observed variation for this region in the B3 domains used in this study **(Fig. 1B and 1C: β4, β8 and β9; Fig. 6A)**, this variability did not impede the capacity of most of these B3 domains to functionally complement Mp*ARF1*. This suggests that either: i) the variation does not affect flexibility, or ii) B3 flexibility is not crucial for MpARF1 DNA-binding *in vivo*. In support of the second point, the B3-hinge flexibility was hypothesized based on DNA-binding affinities to IR motifs (Boer *et al.*, 2014), which are motifs that do not seem to be bound by MpARF1 *in vivo* or DAP-seq (Chapter 4). Thus, whether variation in B3-hinge flexibility is a crucial characteristic of ARF functional divergence remains to be experimentally proven. Since MpARF1 could bind an IR7 motif *in vitro* (Kato *et al.*, 2020), the B3-chimeras created in this chapter could be used to investigate the contributions of variation in the B3-hinge domain.

To conclude, our results highlight that the DNA-binding preferences of the A and B-ARF B3 domains have been acquired in the last common ancestor – the proto-A/B-ARF – and have been retained to similar capacities in all streptophyte lineages. Additionally, these ancestral capacities have been retained in B3 domains of the expanded A-ARF families of angiosperms, in addition to potential neofunctionalization or sub functionalization. For the C-ARF B3 domains in contrast, although previous results indicated a capacity to bind similar AuxRE sites as A and B-ARFs, our results indicate their DNA-binding preferences have diverged.

Material & Methods

Plasmid construction

Gateway entry vectors (pENTr221) containing full-length Mp*ARF1* coding sequence (M1-P925) without stop codon (renamed *pJR86_ENTr_MpARF1CdsNo**), and Gateway destination vector *pHKDW038* ($_{pro}$MpARF1:LR-Citrine, cloned into pMpGWB301; Ishizaki *et al.*, 2015) were kindly provided by Hirotaka Kato. The *proMpARF1* covers 4869 bp upstream the TSS of Mp*ARF1.* For the B3 swaps, we exchanged the endogenous Mp*ARF1* B3 by the target B3 at the exact residues shown in **Fig. 1C**. We PCR amplified *pJR86* with primers JR436 and JR437 (for B3 swaps with Mp*ARFs*, *AtARFs* and *CaARF*), or JHG296 and JHG297 (for *SmARFs* and *PmARFs*; see Supplementary table 1) to remove the B3 domain of Mp*ARF1*, and PCR amplified the other B3 domains using primers with overhangs homologous to *pJR86* (see **Supplementary table 1**). Amplicons were treated with *DpnI* (FD1703, Thermo Scientific) and used in a HiFi Assembly (E2621, New England Biolabs). Correct clones were verified by Sanger sequencing (Integrated DNA Technologies) the entire Mp*ARF1-B3*^{chimera} CDS. The CDS was subcloned to *pHKDW038* using Gateway LR Clonase II (11791020, Invitrogen) and correct orientation and in-frame fusion to *Citrine* verified by Sanger sequencing the fusion site. B3 domains of Mp*ARF2* and Mp*ARF3* were subcloned from plasmid templates used in (Kato *et al.*, 2020), *AtARF5*, *6, 7, 8* and *19* from Arabidopsis leaf or flower cDNA, *CaARF* and *PmARFs* from *C. atmophyticus* and *P. margaritaceaum* cDNA respectively, and *SmARF3* and *SmARF4* from *E.coli* codon-optimized gBlocks with JGH298 and JGH299 (Integrated DNA Technologies).

Plant materials and culture conditions

The male accession of *M. polymorpha*, Takaragaike-1 (Tak-1), the null mutant Mp*arf1-4* in Tak-1 background (Kato *et al.*, 2017) and generated mutant lines were maintained axenic and asexually. Gemmae or sectioned apical notches (~ 10 mm^2) were propagated on half-strength Gamborg's B5 medium (hereafter ½ B5; G0209, Duchefa Biochemie) containing 1% agar (hereafter solid ½ B5) under 50-60 µmol photons m^{-2} s^{-1} white fluorescent light at 22°C under long day (16h light, 8h dark) conditions.

Transformation of Mp*arf1-4*

Mp*arf1-4* thallus material was transformed with *pHKDW038* carrying Mp*ARF1-B3*^{chimera} using an *Agrobacterium tumefaciens* protocol modified from (Kubota *et al.*, 2013). Mp*arf1-4* thallus material was grown anywhere between 2 weeks and 3 months (before the onset of dying tissue) prior to transformation. Green thallus material was cut indiscriminately into tiny pieces (< 1 mm^2) using a broad razor blade and added into 19 ml of liquid 0M51C medium. To this, a 3 ml overnight culture of *A. tumefaciens*, pelleted and resuspended in 1 ml 0M51C, was added and incubated for 1-2 days at 22°C under

long day conditions. Thalli pieces were washed by centrifuging and decanting the 0M51C medium, followed by three washing steps with 30 ml of sterile water with 0.005% Triton-X100. At each washing step, thalli pieces were vortexed thoroughly for 5 min. After the last wash, thalli pieces were incubated in 10 ml water with 0.005% Triton-X100 and 1 mg/ml cefotaxime for 1-2 hours on an orbital shaker. Thalli pieces were grown on solid ½ B5, with 100 µg/ml cefotaxime, 10 mg/ml hygromycin (Mp*arf1-4* background) and 0.5 µM chlorsulfuron (*pHKDW038*). Non-chimeric lines (G_1 generation) were established from T_1 lines via single gemmae. Gemmae from G_1 plants (G_2 generation) or later generations, from multiple independent T_1 lines were used for experiments.

Phenotyping experiments

Fourteen-day growth experiments with or without NAA were performed by growing gemmae on solid ½ B5 with 3 µM NAA or adding an equivalent volume of DMSO at 50-60 µmol photons m^{-2} s^{-1} continuous white fluorescent light at 22°C. After 14 days, pictures were taken with a Canon EOS250D camera with EF-S 18-55 IS STM lens.

MpARF1 structure

The predicted full length MpARF1 structure was generated in Chapter 3 using AlphaFold2 (Jumper *et al.*, 2021; Varadi *et al.*, 2021). Structural figures were created with PyMOL Molecular Graphics System (version 2.5.5, Schrödinger, LLC). Electrostatic interactions were determined with the *distance* command with *cutoff = 3.2* and *mode = 2*. Calculating conservation scores of G234 and surrounding residues was done with the ConSurf webserver (Ashkenazy *et al.*, 2016) on the crystal structure of AtARF1 (PDB: 6ycq) as described in Chapter 3.

Supplementals

Supplemental table 1 | List of primers used in this study

PrimerID	Sequence (5'-3')
JR436_pJR86_R237_F	gctaacaggcaacagacag
JR437_pJR86_D132_R	Tgtaggttgtttggtgtccctt
JR438_MpARF2-R164_F	gggacaccaaacaacctacaCGCAGTTTTACCAAAACCCTC
JR439_MpARF2-R266_R	gctgtctgttgcctgttagcACGCCGGACTCCTACTCTG
JR440_MpARF3-A158_F	gggacaccaaacaacctacaGCCTCGTTCGCCAAGACGCTC
JR441_MpARF2-R260_R	gctgtctgttgcctgttagcCCTGCGCACGCCGACGCA
JR442_AtARF5-E156_F	gggacaccaaacaacctacaGAGTTTTTCTGCAAAACACTTACTG
JR443_AtARF5-R258_R	gctgtctgttgcctgttagcACGCCTAACACCGACCATAAG
JR444_AtARF6-N127_F	gggacaccaaacaacctacaAACTATTTCTGTAAAACTCTGAC
JR445_AtARF6-R229_R	gctgtctgttgcctgttagcTCGTCTTATACCAAGAAGTAATTG
JR446_AtARF7-E125_F	gggacaccaaacaacctacaGAATTTTTCTGCAAAACCC
JR447_AtARF7-R227_R	gctgtctgttgcctgttagcACGTCTTATCCCCAACAG
JR448_AtARF8-N124_F	gggacaccaaacaacctacaAATTATTTTTGTAAGACTCTCAC
JR449_AtARF8-H226_R	gctgtctgttgcctgttagcATGACGAATTCCCAAAAAG
JR450_AtARF19-E124_F	gggacaccaaacaacctacaGAGTTTTTTTGCAAGACTC
JR451_AtARF19-R226_R	gctgtctgttgcctgttagcACGTCTTATACCCAACATC
JR454_CaARF-V105_F	gggacaccaaacaacctacaGTGTCGATAGCAAAGGTGCTC
JR455_CaARF-R209_R	gctgtctgttgcctgttagcGCGCCGGATGCCTAGC
JHG296_pJR86_N239_F	AACAGGCAACAGACAGCAATG
JHG297_pJR86_D132_R	TGTAGGTTGTTTGGTGTCCC
JHG298_SmARF_F	GACACCAAACAACCTACAGTGAGCTTCTGCAAGATG
JHG299_SmARF_R	TTGCTGTCTGTTGCCTGTTAGTTCTCTTAATGCCAAG
JHG300_PmARF3_F	GACACCAAACAACCTACAACTGCCTACTTCAAAGTCC
JHG301_PmARF3_R	TTGCTGTCTGTTGCCTGTTTGCTCTTCGAACACCAATC
JHG302_PmARF4_F	GACACCAAACAACCTACAGAGCACATCGCCAGCAAGACG
JHG303_PmARF4_R	TTGCTGTCTGTTGCCTGTTTGCACGTTGTACTGCCAGC

Chapter

7 General Discussion

The *AUXIN RESPONSE FACTORS* (*ARFs*) family of transcription factors (TFs) are central to the nuclear auxin pathway (NAP) and involved in virtually all aspects of the growth and development of plants (Berleth and Jurgens, 1993; Bowman *et al.*, 2017; Flores-Sandoval *et al.*, 2015; Flores-Sandoval *et al.*, 2018a; Goetz *et al.*, 2006; Hardtke *et al.*, 2004; Kato *et al.*, 2015; Kato *et al.*, 2017; Nagpal *et al.*, 2005; Okushima *et al.*, 2007; Okushima *et al.*, 2005; Tabata *et al.*, 2010; Weijers and Wagner, 2016; Wilmoth *et al.*, 2005; Wu *et al.*, 2006). A proto-ARF originated in the streptophyte algae that duplicated to form A/B and C-ARFs that can be identified in extant streptophyte algae, and duplicated again to form the three distinct ARF families in the last common ancestor (LCA) of land plants: A-, B- and C-ARFs. Following land plant evolution and whole genome duplications, the A-, B- and C-ARF separated into subfamilies in the gymnosperms and angiosperms respectively (Bowman *et al.*, 2021; Carrillo-Carrasco *et al.*, 2023; Flores-Sandoval *et al.*, 2018a; Kato *et al.*, 2018; Martin-Arevalillo *et al.*, 2019; Mutte *et al.*, 2018). ARFs transduce the auxin signalling dynamics into a transcriptional response and as such, the key to their function are their DNA-binding preferences towards auxin response elements (AuxRE; (Kato *et al.*, 2015; Kato *et al.*, 2020; Tiwari *et al.*, 2003; Ulmasov *et al.*, 1999a; Weijers and Wagner, 2016). The ARF DNA-binding domain (DBD) consists of a B3 domain that binds AuxRE, and a dimerization domain that facilitates binding to bipartite AuxRE sites (Boer *et al.*, 2014). Angiosperm ARFs exhibit family-specific differences in their bipartite AuxRE spacing preferences (Boer *et al.*, 2014; Galli *et al.*, 2018; Martin-Arevalillo *et al.*, 2019; Stigliani *et al.*, 2019). Although the ARFs are widespread in the streptophyte lineages, their DNA-binding preferences have primarily been investigated in angiosperm species. Therefore, understanding the DNA-binding preferences of ARFs in species other than angiosperms, elucidating the evolutionary conservation of these preferences and the evolvability thereof are crucial in understanding the evolution of ARFs. In this thesis, we used a combination of approaches to investigate the evolvability and conservation of the ARF-AuxRE interface.

The minimalistic nuclear auxin pathway of *Marchantia polymorpha*

M. polymorpha represents the least complex NAP identified to date and represents the ancestral state of the last common embryophyte ancestor. Previous studies suggested all auxin signalling transduction occurred primarily through the A-class Mp*ARF1*, but tested only a limited number of genes and could therefore not indicate whether Mp*ARF1* is the sole auxin-responsive TF in *M. polymorpha* (Flores-Sandoval *et al.*, 2018a; Kato *et al.*, 2017; Kato *et al.*, 2020; Mutte *et al.*, 2018). In **Chapter 4** and **Chapter 5**, we demonstrated that the Mp*arf1-4* transcriptome was unresponsive to auxin, affirming that Mp*ARF1* is the sole TF transducing the auxin signal in *M. polymorpha*. The B-class Mp*ARF2* and C-class Mp*ARF3* are therefore not involved in auxin transducing directly, supporting previous

studies (Flores-Sandoval *et al.*, 2015; Flores-Sandoval *et al.*, 2018a; Flores-Sandoval *et al.*, 2018b; Kato *et al.*, 2020). MpARF2 was previously shown to act as an attenuator of MpARF1 activity, independent of auxin signalling, by competing for similar DNA sites (Kato *et al.*, 2020). DNA-binding domain (DBD) swap experiments indicated Mp*ARF2* could only partially functionally complement for Mp*ARF1 in vivo*, suggesting that their DNA-binding preferences only partially match (Kato *et al.*, 2020). Because ARF DNA-binding is comprised of two aspects – binding AuxRE and spacing preferences – the specific differences between Mp*ARF1* and Mp*ARF2* remained unclear. In **Chapter 6**, we uncoupled these aspects by swapping the Mp*ARF1-B3* domain with the B3 of Mp*ARF2* and found that both B3 domains are functionally similar. These results suggest that the B3 domains of both ARFs share similar AuxRE binding preferences *in vivo*, implying that differences in spacing preferences of the MpARF2 DBD exclude binding to the full set of MpARF1 target genes. Notably, *in vitro* DNA-binding studies with maize class A- and B-ARFs were inconclusive on whether these ARFs bind the exact same DNA sites (i.e., competition for DNA-binding) or in close proximity (co-regulation) (Galli *et al.*, 2018). The exact mechanism how Mp*ARF2* attenuates Mp*ARF1* therefore requires resolving the DNA-binding preferences of MpARF2 to compare them with the preferences of MpARF1 that we resolved in **Chapter 4**. Collectively, this work supports the hypothesis that Mp*ARF1* is the sole auxin transducing TF in *M. polymorpha* and that MpARF1 and MpARF2 have different spacing preferences.

Conservation of the ARF DBD structure

The crystal structures of AtARF1, AtARF5 and MpARF2 revealed that the ARF DBD structures are largely similar (Boer *et al.*, 2014; Freire-Rios *et al.*, 2020; Kato *et al.*, 2020), which indicates that the overall DBD structure was conserved for 500 million of years of independent evolution between the A and B-class ARFs, as well as between Arabidopsis and *M. polymorpha*. This aligned with the broader observation that protein structure and function tend to be conserved, even when the protein sequences diverge (Choi and Kim, 2006; Flores *et al.*, 1993; Illergård *et al.*, 2009; Yu *et al.*, 2023). This trend is exemplified by the B3 superfamily of plant-specific transcription factors, comprising the LAV, RAV, ARF and REM families (Swaminathan *et al.*, 2008). These B3 families share a similar open β-barrel structure, despite low sequence similarity and diverse DNA-binding capabilities (King *et al.*, 2013; Sasnauskas *et al.*, 2018a; Sasnauskas *et al.*, 2018b; Wu *et al.*, 2018; Yamasaki *et al.*, 2004). However, outside of a few homology models (Kato *et al.*, 2020; Mutte *et al.*, 2018), no larger scale comparisons had been made between sequence conservation and structural conservation. We therefore investigated the ARF DBD protein sequence and structural fold across streptophyte lineages in **Chapter 3**. We observed that despite sequence identities as low as 30%, the DBDs' structural folds were strikingly similar between A, B and C-class ARFs, likely due to the higher conservation of

buried, structural residues, as opposed to the variable exposed residues. Furthermore, we noted C-class-specific and species-specific altered topologies. The C-class ARFs, while generally aligning with the overall DBD structure, had an extended loop that connects the B3 domain with the rest of the DBD as compared to A and B-class ARFs, as noted previously (Mutte *et al.*, 2018; Martin-Arevalillo *et al.*, 2019). Furthermore, the A/B-ARFs from *Penium margaritaceaum* (Zygnematophyceae) exhibit notable deviations in their predicted DBD folds, including variation in the dimerization helix's angles and lengths, as predicted by both AlpaFold2 and EMSFold (Jumper *et al.*, 2021; Lin *et al.*, 2023; Varadi *et al.*, 2021). Determining whether these unique topologies are prediction artefacts or represent biologically relevant adaptations requires the analysis of more predicted structures and solving a PmARF crystal structure to serve as a benchmark. Nonetheless, we found that the DBD fold is an ancient protein structure that was established in the LCA of all ARFs in the streptophyte algae and remains overall similar in extant ARFs.

Although sequence-based phylogenetic trees could pinpoint the origin of the full ARF-DBD to the streptophyte algae (Bowman *et al.*, 2021; Carrillo-Carrasco *et al.*, 2023; Flores-Sandoval *et al.*, 2018a; Kato *et al.*, 2018; Martin-Arevalillo *et al.*, 2019; Mutte *et al.*, 2018), it remains a question how the ARFs evolved in the first place. This question, as well as the evolution of alternative topologies with possible biological relevance (as displayed by the PmARFs), can only be resolved by increasing the sample size of structural homologs. Within the ARFs, the AuxRE-interacting B3 domain is embedded within the larger DBD structure (Boer *et al.*, 2014). Although the B3 domain superfamily of TF DNA-binding domains likely originated within the chlorophytes (Mutte *et al.*, 2018; Swaminathan *et al.*, 2008), the B3 fold is structurally similar to prokaryotic restriction enzymes (Swaminathan *et al.*, 2008; Yamasaki *et al.*, 2004). The ARF DBD further comprises a dimerization domain (DD) and ancillary domain (AD), both of which are found as subdomains in non-ARF proteins of red algae, a clade that do not have any ARFs (Mutte *et al.*, 2018). However, sequence-only phylogeny may miss remote structural homology, such as the B3 domains structural similarities to restriction enzymes (Yamasaki *et al.*, 2004), and may therefore fail to trace the full evolutionary history. Since our investigation in **Chapter 3**, AlphaFold2 has been used to predict the protein structures of all proteins deposited in the UniProt database. This has led to advancements such as hierarchical clustering of the entire structurome, the discovery of novel protein folds and functions, the classification of unknown genes, and the implementation of structure-based, deep phylogenies of protein families (Barrio-Hernandez *et al.*, 2023; Durairaj *et al.*, 2023; Moi *et al.*, 2023; Wu *et al.*, 2023; Yu *et al.*, 2023). Specifically for the ARFs, such advances could supplement recent deep phylogenetic studies by exploring remote homologies of the AD and DD (Carrillo-Carrasco *et al.*, 2023; Mutte *et al.*, 2018), aid in the identification of structural features that distinguish A, B and C-ARFs, or uncover species-specific structural novelties like the unique topologies of the PmARFs.

Conservation of A and B-ARF AuxRE binding preferences

The AuxRE preferences of the ARF-B3 domain have previously been investigated in detail only for A- and B-ARFs from the angiosperms Arabidopsis and maize, using *in vitro* methods (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014; Galli *et al.*, 2018; O'Malley *et al.*, 2016; Stigliani *et al.*, 2019). Therefore, our understanding of AuxRE preferences outside of the angiosperms has been limited. AuxRE are hexamers of nucleotides TGTCNN, where TGTC forms the immutable core sequence and N5 and N6 can be variable (Boer *et al.*, 2014; Franco-Zorrilla *et al.*, 2014; Ulmasov *et al.*, 1999b; Ulmasov *et al.*, 1997b). A- and B-ARFs from angiosperms have the highest affinity to TGTCGG elements *in vitro* and *in vivo* (Boer *et al.*, 2014; Galli *et al.*, 2018; Liao *et al.*, 2015; Stigliani *et al.*, 2019), suggesting these preferences were established in an ancestral A/B-ARF. We confirmed this hypothesis in **Chapter 4**, by identifying TGTCGG elements as the preferred elements of MpARF1 *in vitro*.

Within an expanded family of TFs, similar preferences for high affinity DNA elements does not guarantee functional equivalence. For example, DBD swaps between the *Drosophila melongaster Hox* family of TFs indicated a functional divergence between the members, despite them sharing the same preferences towards high affinity DNA elements (Crocker *et al.*, 2015; Kribelbauer *et al.*, 2019). Instead, the members subspecialized by acquiring unique preferences towards medium affinity sites (Crocker *et al.*, 2015; Kribelbauer *et al.*, 2019). In Arabidopsis, both AtARF1 (B-ARF) and AtARF5 (A-ARF) could bind a variety of TGTCNN elements *in vitro*, but only seemed to differ in general affinity towards all TGTCNN elements rather than subspecialization towards different elements. This led to the hypothesis that all A- and B-ARF B3 domains could be functionally equivalent, which we assessed in **Chapter 6**, via B3 swap complementation assays between MpARF1 and B3 domains of the *M. polymorpha* B-ARF (Mp*ARF2*), Arabidopsis A-ARFs (*AtARF5, 6, 7, 8,* and *19*), and A/B-ARFs from two streptophyte algae. We observed that the B3 domains from all these lineages were functionally equivalent to MpARF1, indicating these B3 domains share a conserved, ancient DNA-binding preference. However, based on this experiment we cannot exclude the possibility that the A-ARFs further subspecialized in angiosperms, like the *Hox* genes (Crocker *et al.*, 2015), which will need to be addressed in Arabidopsis itself (see **Chapter 6**). We therefore find evidence that the TGTCNN binding preferences were already established in the LCA of all A-, B- and A/B-ARFs and that these ancestral preferences have been maintained in extant streptophyte lineages, including in extended A-ARF families of angiosperms.

The conservation of DNA-binding preferences by A-, B- and A/B-ARFs could be attributed to the strong conservation of the DNA-binding residues, as explored in **Chapter 3**. These results suggest i) that strong evolutionary pressure maintains DNA-binding as is, and/or ii) that mutations on DNA-binding residues have dramatic effects on affinity. We investigated these scenarios in **Chapter 5**, where we identified that

variation on DNA-binding residues was rare, occurred only in species with expanded ARF families, and possibly only in pseudogenes. That last option was not explored further in our work. Variation for the B3 domain residues binding to the TGTC-core sequence (i.e., MpARF1-R194, P197 and R199) all strongly impaired DNA-binding of MpARF1, supporting the hypothesis that ARF-AuxRE binding is highly stringent. In **Chapter 3**, we found evolutionary couplings for two of these residues (P197 and R199) and it will be interesting to explore the origins of these couplings and how they could affect DNA-binding. Surprisingly, in **Chapter 5** we identified two residues – MpARF1-H146 and Q196 – on which variation seemingly did not affect MpARF1 function, despite their high conservation across the streptophytes (see **Chapter 3**, **Chapter 5** and **Chapter 6**). Yet most MpARF1-H146 and Q196 mutants did not impair function in our assays. This suggests H146 and Q196 might be important residues for MpARF1 function in growth aspects not assessed (e.g, gemmae production), or when grown outside of controlled conditions. Since field conditions may trigger novel molecular interactions compared with controlled conditions in Arabidopsis (He *et al.*, 2014; Sterken *et al.*, 2023), it will be interesting to test these H146 mutants in conditions mimicking their natural habitat or under stress conditions. Furthermore, these mutants (H146 in particular due to its importance in binding N5 and N6 of TGTCNN) could have subtle effects on the specificity of TGTCNN elements, which needs to be validated by DNA-binding assays. To conclude, we therefore found that the ARF-AuxRE interface is strongly conserved, likely because mutations on the DNA-binding residues experience purifying selection since they are so detrimental to proper function.

Variability in A and B-ARF spacing preferences

ARF DNA-binding is composed of two aspects: binding to AuxRE and spacing between two AuxRE, and whether ARF spacing preferences are conserved between different streptophyte lineages was unknown. As discussed above, since AuxRE-binding is similar between the A- and B-ARFs, their main difference lies in their spacing preferences. These preferences have been characterized for angiosperms ARFs, and briefly: A-ARFs can bind direct repeat motifs with 5 basepair spacing (DR5) and inverted repeat motifs with 4 to 9 basepairs of spacing (IR4-9), whereas B-ARFs are limited to binding IR7/8 motifs (for a more comprehensive overview, see Chapter 2). In **Chapter 3**, we noted that the residues of the canonical dimerization interface, as originally identified by Boer *et al.* (2014) on IR motifs, were not conserved across the plant kingdom. We interpreted this to indicate that the interface was highly variable and could lead to alternative spacing preferences for ARFs. To investigate whether A-ARF spacing preferences were conserved, we characterized the spacing preferences of MpARF1 and performed promoter enrichment analysis of *M. polymorpha* auxin-responsive genes in **Chapter 4**. Strikingly, we found that MpARF1 does not bind DR5 and IR4-9 motifs like the angiosperm A-ARFs, but instead

prefers DR[-2] and DR2. Indeed, this could explain why the widely used DR5 reporter is not auxin-responsive in *M. polymorpha* (Ishizaki *et al.*, 2012). It therefore indicates that ARF spacing preferences have diverged between the liverworts and angiosperms.

Modelling the binding of MpARF1 to DR[-2] and DR2 suggest they are bound by monomers and dimers respectively (see **Chapter 4**). Our structural modelling indicates that for two monomers to bind to DR[-2] it would require the displacement of a DNA-binding loop for one of the ARF-DBD monomers. As we did not identify potential dimerization interfaces, it is likely that the DR[-2] site is bound by a monomeric MpARF1. Modelling studies suggests eukaryotic TFs non-specifically bind DNA with intrinsically disordered regions (IDR), like the ARF middle region (Roosjen *et al.*, 2018), to facilitate 'sliding' along the DNA to search for their cognate DNA sites (Vuzman *et al.*, 2010a; Vuzman *et al.*, 2010b). In this scenario, the DR[-2] site could improve the chance of a sliding ARF to stall and thus increase affinity, a hypothesis that requires biophysical validation. Modelling suggests the DR2 motif could be bound by homodimers via a novel interaction interface, potentially cooperatively. This interface, cooperativity, as well as binding to DR[-2] and DR2 motifs require functional validation *in vivo*.

Differences in spacing preferences within a single TF family, A-ARFs in this case, is likely extremely rare, given that the vast majority of 242 *Drosophila melongaster* (fruit fly) TFs, including homodimeric TFs, have strikingly similar DNA-binding specificities as their human and mice orthologs, even after 600 million years of separate evolution (Jolma *et al.*, 2013; Nitta *et al.*, 2015). Even if rare, divergent DNA-specificities in orthologous TFs could occur if the ancestral TF exhibited promiscuous DNA-binding. This is exemplified by the TF *LEAFY* (*LFY*), which is conserved as a single copy in most streptophyte species (Gao *et al.*, 2019; Sayou *et al.*, 2014). LFY from different streptophyte species subspecialized for one of three DNA-binding preferences (mainly differences in flanking nucleotides and spacing between two sites): type I, II and III, of which type I LFY cannot bind type II and type III, and so forth. However, the authors identified a promiscuous LFY that exhibited all three types of DNA-binding as the likely ancestral state, which subsequently subspecialized within the streptophyte lineages (Gao *et al.*, 2019; Sayou *et al.*, 2014). For the A-ARFs, a promiscuous ancestor could explain the observed spacing differences. For instance, some maize A-ARFs could bind DR2 motifs *in vitro*, despite DR2 motifs not being enriched in auxin-induced gene promoters (although a weak enrichment of DR2 motifs in auxin-repressed gene promoters was found in Arabidopsis; Freire-Rios *et al.*, 2020). Vice versa, although we did not identify DR5 or IR7 motifs (see **Chapter 4**), MpARF1 and MpARF2 are capable of binding IR7 in single-molecule *in vitro* DNA-binding assays (Kato *et al.*, 2020) and MpARF1 can bind DR5 motifs in a heterologous system (see **Chapter 4**). To investigate the possibility of a promiscuous ancestor for the A-ARFs, sampling the spacing preferences of ARFs from different streptophyte lineages would be necessary, particularly from other bryophytes like the moss *Physcomitrella patens*, and

tracheophytes (lycophytes and ferns).

C-ARFs have diverged from A/B-ARFs in their DNA-binding

Previous studies have suggested that the *M. polymorpha* C-ARF (MpARF3) has little functional overlap with MpARF1 and MpARF2. Domain swaps indicate that full-length MpARF3 and MpARF3-DBD cannot functionally replace MpARF1 (Kato *et al.*, 2020). This is likely explained by the spacing preferences of C-ARFs, as *in vitro* DNA-binding assays suggest that C-ARFs have different spacing preferences compared to A- and B-ARFs (Martin-Arevalillo *et al.*, 2019). In **Chapter 6**, we used B3 swaps with MpARF3 to investigate whether AuxRE preferences were conserved between MpARF1 and MpARF3. We found that the MpARF3 B3-domain could only partially functionally replace MpARF1, suggesting their AuxRE preferences only partially overlap. In **Chapter 3** and **Chapter 5**, we identified that three of the TGTC-core-binding residues are conserved between A-, B- and C-ARFs (i.e, MpARF1-R194, P197 and R199), while the remaining two were substituted with C-ARF-specific, conserved residues: H146 with a glycine (G), and Q196 with a threonine (T). This conservation suggest that the MpARF3 B3-domain could likely bind TGTCNN-based elements, which has been confirmed *in vitro* for other C-ARFs (Martin-Arevalillo *et al.*, 2019). However, the replacement of H146 with a glycine might cause the MpARF3-B3 to preferentially bind TGTCNN elements that are different from A- and B-ARFs, e.g. TGTCGG may not be the highest affinity element for C-ARFs. Indeed, in **Chapter 5**, we observed that the MpARF1-H146G variant complemented Mp*arf1-4* in all assessed aspects, but produced less gemma cups, suggesting that gemma-cup-related MpARF1 target genes contain TGTCNN elements that are bound differently by the MpARF1-H146G variant. Since the variant was still capable of restoring most phenotypes of the Mp*arf1-4* mutant, it suggests more mutations must have occurred to diverge the A/B-ARFs from the C-ARFs. We therefore inferred that the C-ARFs, in addition to their spacing preferences and on other domains (Kato *et al.*, 2020; Martin-Arevalillo *et al.*, 2019), exhibit different DNA-binding preferences as the A- and B-ARFs.

From minimal to expanded systems

In comparison to *M. polymorpha*, the ARF families have become more expanded in Arabidopsis. In *M. polymorpha*, the single A-ARF is the sole auxin transducing TF (see **Chapter 4** and **Chapter 5**) and therefore must be involved in all NAP-mediated developmental processes. In Arabidopsis however, the A-ARFs gained subspecialized roles as evidenced by their unique spatial expression profiles and the distinct phenotypic effects of their mutants (Berleth and Jurgens, 1993; Goetz *et al.*, 2006; Hardtke *et al.*, 2004; Nagpal *et al.*, 2005; Okushima *et al.*, 2007; Okushima *et al.*, 2005; Rademacher *et*

al., 2011; Tabata *et al.*, 2010; Weijers and Wagner, 2016; Wilmoth *et al.*, 2005; Wu *et al.*, 2006). To investigate whether the A-ARFs retained an ancestral DNA-binding preference, we performed B3-swaps with the ortholog MpARF1 in **Chapter 6**. Since all Arabidopsis A-ARF B3-domains were able to functionally replace the B3 of MpARF1, we confirm that these ARFs must retain an ancestral DNA-binding specificity similar to MpARF1. As stated above, we cannot exclude the possibility that A-ARFs may have acquired novel DNA-specificities in addition to retaining the ancestral binding. The conservation of ancestral DNA-binding preferences of ARFs matches the finding for animal TFs, wherein the vast majority of TFs retain the same DNA-binding preferences, as referenced above (Jolma *et al.*, 2013; Nitta *et al.*, 2015). It is therefore likely that subspecialisation must have occurred on different aspects or different domains of these ARFs.

There are several ways how TFs can acquire novel functions after gene duplications. First, the unique mutant phenotypes and expression patterns of the A-ARFs suggests subspecialization (Okushima *et al.*, 2005; Rademacher *et al.* 2011). Whether this is attributable to differences in spacing preferences could be tested by performing DBD swaps between the A-ARFs. Second, expansion of a gene family could lead to novel intrafamilial interactions. The heterodimerizing MADS-box family of TFs, for example, is associated with the development of flowers, as the family largely expanded in the ancestor of angiosperms, established novel interaction networks and expression domains, despite having highly similar DNA-binding preferences (Aerts *et al.*, 2018; Amborella Genome, 2013; Zhang *et al.*, 2018). The expansion of the ARF family and subsequent partially overlapping expression domains could have similarly led to novel interactions. However, the biological relevance of ARF heterodimerization has not been investigated. Third, domains outside of the DBDs might have acquired novel functions or interactions. For example, the MYB *DUO POLLEN1* (*DUO1*) is involved in sperm cell differentiation in both *M. polymorpha* and Arabidopsis. In cross-species complementation, full-length Mp*DUO1* was unable to fully complement *Atduo1* mutant, and vice versa. However, swapping only the DBDs led to full complementation, indicating that the domains outside of the DBD acquired species-specific functions (Higo *et al.*, 2018). Indeed, the interactions between A-ARFs and Aux/IAA proteins, which are mediated by the Phox and Bem1 (PB1) domains of ARFs, became highly complex in Arabidopsis and suggest subspecialization for the PB1 domains (Li *et al.*, 2011; Weijers *et al.*, 2005). Domain swapping experiments between the Arabidopsis ARFs will likely lead to more insights into how this family subspecialized.

Conclusion

This thesis investigated the evolution and evolvability of ARF DNA-binding preferences. We identified that the B3 domain residues are highly conserved, including the DNA-binding residues. Although variation for these residues exist within the plant kingdom, they are generally detrimental to proper function of the ARFs. This explains how the B3-domain of A-, B- and A/B-ARFs derived from plants from different streptophyte lineages can functionally replace the B3-domain of MpARF1. The C-ARF B3 domains are functionally divergent from the A-, B- and A/B-ARFs, likely by having alternative preferences for AuxRE-related DNA elements. MpARF1 was identified as the sole auxin-transducing transcription factor in *Marchantia polymorpha* and shown to have similar AuxRE binding preferences as angiosperm ARFs. Importantly, MpARF1 was revealed to exhibit spacing preferences between two AuxRE that are different from angiosperm ARFs. This work has shown that the AuxRE binding of the ARFs is highly conserved and that ARF subfunctionalization has likely occurred on spacing differences. Future complementation assays and domain swaps in Arabidopsis ARF mutant and *M. polymorpha* MpARF3 mutant backgrounds will further reveal the evolution of the DNA-binding preferences.

References

Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, Rohs R, Mann RS. 2015. Deconvolving the recognition of DNA shape from sequence. *Cell* **161**, 307-318.

Adam K, Gyorgypal Z, Hegedus Z. 2019. DNA Readout Viewer (DRV): visualization of specificity determining patterns of protein-binding DNA segments. *Bioinformatics* **36**, 2286-2287.

Adhav VA, Saikrishnan K. 2023. The Realm of Unconventional Noncovalent Interactions in Proteins: Their Significance in Structure and Function. *ACS Omega* **8**, 22268-22284.

Aditham AK, Markin CJ, Mokhtari DA, DelRosso N, Fordyce PM. 2021. High-Throughput Affinity Measurements of Transcription Factor and DNA Mutations Reveal Affinity and Specificity Determinants. *Cell Syst* **12**, 112-127 e111.

Aerts N, de Bruijn S, van Mourik H, Angenent GC, van Dijk ADJ. 2018. Comparative analysis of binding patterns of MADS-domain proteins in Arabidopsis thaliana. *BMC Plant Biology* **18**, 131.

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33**, 831.

Amborella Genome P. 2013. The Amborella genome and the evolution of flowering plants. *Science* **342**, 1241089.

Amoutzias GD, Robertson DL, Van de Peer Y, Oliver SG. 2008. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci* **33**, 220-229.

Anderson DW, McKeown AN, Thornton JW. 2015. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *Elife* **4**, e07864.

Andrews G, Fan K, Pratt HE, Phalke N, Zoonomia Consortium section s, Karlsson EK, Lindblad-Toh K, Gazal S, Moore JE, Weng Z. 2023. Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science* **380**, eabn7930.

Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N. 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research* **44**, W344-350.

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723.

Baek M, McHugh R, Anishchenko I, Baker D, DiMaio F. 2022. Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. *bioRxiv* 10.1101/2022.09.09.507333.

Bailey TL. 2021. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* **37**, 2834-2840.

Baker CR, Tuch BB, Johnson AD. 2011. Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 7493-7498.

Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M. 2023. Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637-645.

Bartlett A, O'Malley RC, Huang S-sC, Galli M, Nery JR, Gallavotti A, Ecker JR. 2017. Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nature Protocols* **12**, 1659-1672.

Berg J, Willmann S, Lassig M. 2004. Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology* **4**, 42.

Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, Schiegg M, Ales J, Beier T, Rudy M, Eren K, Cervantes JI, Xu B, Beuttenmueller F, Wolny A, Zhang C, Koethe U, Hamprecht FA, Kreshuk A. 2019. ilastik: interactive machine learning for (bio)image analysis. *Nature Methods* **16**, 1226-1232.

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, Khalid F, Zhang W, Newburger D, Jaeger SA, Morris QD, Bulyk ML, Hughes TR. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266-1276.

Berleth T, Jurgens G. 1993. The role of the monopteros gene in organising the basal body region of the Arabidopsis embryo. *Development* **118**, 575-587.

Boer DR, Freire-Rios A, van den Berg WA, Saaki T, Manfield IW, Kepinski S, Lopez-Vidrieo I, Franco-Zorrilla JM, de Vries SC, Solano R, Weijers D, Coll M. 2014. Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. *Cell* **156**, 577-589.

Boldyreva LV, Goncharov FP, Demakova OV, Zykova TY, Levitsky VG, Kolesnikov NN, Pindyurin AV, Semeshin VF, Zhimulev IF. 2017. Protein and Genetic Composition of Four Chromatin Types in Drosophila melanogaster Cell Lines. *Current Genomics* **18**, 214-226.

Bowman JL, Flores Sandoval E, Kato H. 2021. On the Evolutionary Origins of Land Plant Auxin Biology. *Cold Spring Harb Perspect Biol* **13**.

Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu S, Ishizaki K, Yamaoka S, Nishihama R, Nakamura Y, Berger F, Adam C, Aki SS, Althoff F, Araki T, Arteaga-Vazquez MA, Balasubrmanian S, Barry K, Bauer D, Boehm CR, Briginshaw L, Caballero-Perez J, Catarino B, Chen F, Chiyoda S, Chovatia M, Davies KM, Delmans M, Demura T, Dierschke

T, Dolan L, Dorantes-Acosta AE, Eklund DM, Florent SN, Flores-Sandoval E, Fujiyama A, Fukuzawa H, Galik B, Grimanelli D, Grimwood J, Grossniklaus U, Hamada T, Haseloff J, Hetherington AJ, Higo A, Hirakawa Y, Hundley HN, Ikeda Y, Inoue K, Inoue SI, Ishida S, Jia Q, Kakita M, Kanazawa T, Kawai Y, Kawashima T, Kennedy M, Kinose K, Kinoshita T, Kohara Y, Koide E, Komatsu K, Kopischke S, Kubo M, Kyozuka J, Lagercrantz U, Lin SS, Lindquist E, Lipzen AM, Lu CW, De Luna E, Martienssen RA, Minamino N, Mizutani M, Mizutani M, Mochizuki N, Monte I, Mosher R, Nagasaki H, Nakagami H, Naramoto S, Nishitani K, Ohtani M, Okamoto T, Okumura M, Phillips J, Pollak B, Reinders A, Rovekamp M, Sano R, Sawa S, Schmid MW, Shirakawa M, Solano R, Spunde A, Suetsugu N, Sugano S, Sugiyama A, Sun R, Suzuki Y, Takenaka M, Takezawa D, Tomogane H, Tsuzuki M, Ueda T, Umeda M, Ward JM, Watanabe Y, Yazaki K, Yokoyama R, Yoshitake Y, Yotsui I, Zachgo S, Schmutz J. 2017. Insights into Land Plant Evolution Garnered from the Marchantia polymorpha Genome. *Cell* **171**, 287-304 e215.

Calderón Villalobos LI, Lee S, De Oliveira C, Ivetac A, Brandt W, Armitage L, Sheard LB, Tan X, Parry G, Mao H, Zheng N, Napier R, Kepinski S, Estelle M. 2012. A combinatorial TIR1/AFB-Aux/IAA co-receptor system for differential sensing of auxin. *Nature Chemical Biology* **8**, 477-485.

Carrillo-Carrasco VP, Hernandez-Garcia J, Mutte SK, Weijers D. 2023. The birth of a giant: evolutionary insights into the origin of auxin responses in plants. *EMBO Journal* **42**, e113018.

Carroll SM, Ortlund EA, Thornton JW. 2011. Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor. *PLoS Genetics* **7**, e1002117.

Causier B, Ashworth M, Guo W, Davies B. 2012. The TOPLESS interactome: a framework for gene repression in Arabidopsis. *Plant Physiology* **158**, 423-438.

Chandrasegaran S, Carroll D. 2015. Origins of Programmable Nucleases for Genome Engineering. *Journal of Molecular Biology* http://dx.doi.org/10.1016/j.jmb.2015.10.014.

Chen X, Li C, Wang H, Guo Z. 2019. WRKY transcription factors: evolution, binding, and action. *Phytopathology Research* **1**.

Cherenkov P, Novikova D, Omelyanchuk N, Levitsky V, Grosse I, Weijers D, Mironova V. 2018. Diversity of cis-regulatory elements associated with auxin response in Arabidopsis thaliana. *Journal of Experimental Botany* **69**, 329-339.

Chiyoda S, Ishizaki K, Kataoka H, Yamato KT, Kohchi T. 2008. Direct transformation of the liverwort Marchantia polymorpha L. by particle bombardment using immature thalli developing from spores. *Plant Cell Reports* **27**, 1467-1473.

Choi I-G, Kim S-H. 2006. Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences* **103**, 14056-14061.

Chun I, Kim HJ, Hong S, Kim Y-G, Kim M-S. 2023. Structural basis of DNA binding by the NAC transcription factor ORE1, a master regulator of plant senescence. *Plant Communications* **4**.

Clarke ND. 1995. Covariation of residues in the homeodomain sequence family. *Protein Science* **4**, 2269-2278.

Cortal A, Martignetti L, Six E, Rausell A. 2021. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nature Biotechnology* **39**, 1095-1102.

Crocker J, Abe N, Rinaldi L, McGregor AP, Frankel N, Wang S, Alsawadi A, Valenti P, Plaza S, Payre F, Mann RS, Stern DL. 2015. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160**, 191-203.

Das S, de Roij M, Bellows S, Kohlen W, Farcot E, Weijers D, Borst JW. 2022. Selective degradation of ARF monomers controls auxin response in Marchantia. *bioRxiv* 10.1101/2022.11.04.515187.

Das, S. 2023. Quantitative dynamics of auxin signaling proteins in Marchantia polymorpha. Doctoral Thesis, Wageningen University and Research, Wageningen.

Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ. 2002. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *Journal of Molecular Biology* **319**, 1097-1113.

Deng C, Wu Y, Lv X, Li J, Liu Y, Du G, Chen J, Liu L. 2022. Refactoring transcription factors for metabolic engineering. *Biotechnology Advances* **57**, 107935.

Deochand DK, Perera IC, Crochet RB, Gilbert NC, Newcomer ME, Grove A. 2016. Histidine switch controlling pH-dependent protein folding and DNA binding in a transcription factor at the core of synthetic network devices. *Mol Biosyst* **12**, 2417-2426.

Dharmasiri N, Dharmasiri S, Weijers D, Lechner E, Yamada M, Hobbie L, Ehrismann JS, Jürgens G, Estelle M. 2005. Plant development is regulated by a family of auxin receptor F box proteins. *Developmental Cell* **9**, 109-119.

Dickinson BC, Leconte AM, Allen B, Esvelt KM, Liu DR. 2013. Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 9007-9012.

Dominguez C, Boelens R, Bonvin AM. 2003. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **125**, 1731-1737.

Donner TJ, Sherr I, Scarpella E. 2009. Regulation of preprocambial cell state acquisition by auxin signaling in Arabidopsis leaves. *Development* **136**, 3235-3246.

Dorrity MW, Cuperus JT, Carlisle JA, Fields S, Queitsch C. 2018. Preferences in a trait decision determined by transcription

factor variants. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E7997-E8006.

Durairaj J, Waterhouse AM, Mets T, Brodiazhenko T, Abdullah M, Studer G, Tauriello G, Akdel M, Andreeva A, Bateman A, Tenson T, Hauryliuk V, Schwede T, Pereira J. 2023. Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646-653.

Eklund DM, Ishizaki K, Flores-Sandoval E, Kikuchi S, Takebayashi Y, Tsukamoto S, Hirakawa Y, Nonomura M, Kato H, Kouno M, Bhalerao RP, Lagercrantz U, Kasahara H, Kohchi T, Bowman JL. 2015. Auxin Produced by the Indole-3-Pyruvic Acid Pathway Regulates Development and Gemmae Dormancy in the Liverwort Marchantia polymorpha. *Plant Cell* **27**, 1650-1669.

Esvelt KM, Carlson JC, Liu DR. 2011. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499-503.

Flores-Sandoval E, Eklund DM, Bowman JL. 2015. A Simple Auxin Transcriptional Response System Regulates Multiple Morphogenetic Processes in the Liverwort Marchantia polymorpha. *PLoS Genetics* **11**, e1005207.

Flores-Sandoval E, Eklund DM, Hong SF, Alvarez JP, Fisher TJ, Lampugnani ER, Golz JF, Vazquez-Lobo A, Dierschke T, Lin SS, Bowman JL. 2018a. Class C ARFs evolved before the origin of land plants and antagonize differentiation and developmental transitions in Marchantia polymorpha. *New Phytologist* **218**, 1612-1630.

Flores-Sandoval E, Romani F, Bowman JL. 2018b. Co-expression and Transcriptome Analysis of Marchantia polymorpha Transcription Factors Supports Class C ARFs as Independent Actors of an Ancient Auxin Regulatory Module. *Front Plant Sci* **9**, 1345.

Flores TP, Orengo CA, Moss DS, Thornton JM. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Science* **2**, 1811-1826.

Fontana M, Ivanovaite S, Lindhoud S, van der Wijk E, Mathwig K, Berg WVD, Weijers D, Hohlbein J. 2022. Probing DNA - Transcription Factor Interactions Using Single-Molecule Fluorescence Detection in Nanofluidic Devices. *Advanced Biology* **6**, e2100953.

Fontana M, Roosjen M, Crespo Garcia I, van den Berg W, Malfois M, Boer R, Weijers D, Hohlbein J. 2023. Cooperative action of separate interaction domains promotes high-affinity DNA binding of Arabidopsis thaliana ARF transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **120**, e2219916120.

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 of the United States of America* **111**, 2367-2372.

Freire-Rios A, Tanaka K, Crespo I, van der Wijk E, Sizentsova Y, Levitsky V, Lindhoud S, Fontana M, Hohlbein J, Boer DR, Mironova V, Weijers D. 2020. Architecture of DNA elements mediating ARF transcription factor binding and auxin-responsive gene expression in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 24557-24566.

Galli M, Khakhar A, Lu Z, Chen Z, Sen S, Joshi T, Nemhauser JL, Schmitz RJ, Gallavotti A. 2018. The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nat Commun* **9**, 4526.

Gao B, Chen M, Li X, Zhang J. 2019. Ancient duplications and grass-specific transposition influenced the evolution of LEAFY transcription factor genes. *Commun Biol* **2**, 237.

Gao L, Cox DBT, Yan WX, Manteiga JC, Schneider MW, Yamano T, Nishimasu H, Nureki O, Crosetto N, Zhang F. 2017. Engineered Cpf1 variants with altered PAM specificities. *Nature Biotechnology* **35**, 789-792.

Garvie CW, Wolberger C. 2001. Recognition of Specific DNA Sequences. *Molecular Cell* **8**, 937-946.

Gera T, Jonas F, More R, Barkai N. 2022. Evolution of binding preferences among whole-genome duplicated transcription factors. *Elife* **11**.

Godoy M, Franco-Zorrilla JM, Perez-Perez J, Oliveros JC, Lorenzo O, Solano R. 2011. Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. *Plant Journal* **66**, 700-711.

Goetz M, Vivian-Smith A, Johnson SD, Koltunow AM. 2006. AUXIN RESPONSE FACTOR8 is a negative regulator of fruit initiation in Arabidopsis. *Plant Cell* **18**, 1873-1886.

Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.

Guilfoyle TJ, Ulmasov T, Hagen G. 1998. The ARF family of transcription factors and their role in plant hormone-responsive transcription. *Cellular and Molecular Life Sciences CMLS* **54**, 619-627.

Hamann T, Benkova E, Baurle I, Kientz M, Jurgens G. 2002. The Arabidopsis BODENLOS gene encodes an auxin response protein inhibiting MONOPTEROS-mediated embryo patterning. *Genes & Development* **16**, 1610-1615.

Hames C, Ptchelkine D, Grimm C, Thevenon E, Moyroud E, Gerard F, Martiel JL, Benlloch R, Parcy F, Muller CW. 2008. Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *EMBO Journal* **27**, 2628-2637.

Han M, Park Y, Kim I, Kim EH, Yu TK, Rhee S, Suh JY. 2014. Structural basis for the auxin-induced transcriptional regulation by Aux/IAA17. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 18613-18618.

Hardtke CS, Berleth T. 1998. The Arabidopsis gene MONOPTEROS encodes a transcription factor mediating embryo axis

formation and vascular development. *EMBO Journal* **17**, 1405-1411.

Hardtke CS, Ckurshumova W, Vidaurre DP, Singh SA, Stamatiou G, Tiwari SB, Hagen G, Guilfoyle TJ, Berleth T. 2004. Overlapping and non-redundant functions of the Arabidopsis auxin response factors MONOPTEROS and NONPHOTOTROPIC HYPOCOTYL 4. *Development* **131**, 1089-1100.

Harris BJ, Clark JW, Schrempf D, Szollosi GJ, Donoghue PCJ, Hetherington AM, Williams TA. 2022. Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. *Nat Ecol Evol* **6**, 1634-1643.

Harteis S, Schneider S. 2014. Making the bend: DNA tertiary structure and protein-DNA interactions. *International Journal of Molecular Sciences* **15**, 12335-12363.

Hashimoto H, Wang D, Steves AN, Jin P, Blumenthal RM, Zhang X, Cheng X. 2016. Distinctive Klf4 mutants determine preference for DNA methylation status. *Nucleic Acids Research* **44**, 10177-10185.

He H, de Souza Vidigal D, Snoek LB, Schnabel S, Nijveen H, Hilhorst H, Bentsink L. 2014. Interaction between parental environment and genotype affects plant and seed performance in Arabidopsis. *Journal of Experimental Botany* **65**, 6603-6615.

Higo A, Kawashima T, Borg M, Zhao M, López-Vidriero I, Sakayama H, Montgomery SA, Sekimoto H, Hackenberg D, Shimamura M, Nishiyama T, Sakakibara K, Tomita Y, Togawa T, Kunimoto K, Osakabe A, Suzuki Y, Yamato KT, Ishizaki K, Nishihama R, Kohchi T, Franco-Zorrilla JM, Twell D, Berger F, Araki T. 2018. Transcription factor DUO1 generated by neo-functionalization is associated with evolution of sperm differentiation in plants. *Nature Communications* **9**.

Higo A, Niwa M, Yamato KT, Yamada L, Sawada H, Sakamoto T, Kurata T, Shirakawa M, Endo M, Shigenobu S, Yamaguchi K, Ishizaki K, Nishihama R, Kohchi T, Araki T. 2016. Transcriptional framework of male gametogenesis in the liverwort Marchantia polymorpha L. *Plant and Cell Physiology* **57**, 325-338.

Hopf TA, Green AG, Schubert B, Mersmann S, Scharfe CPI, Ingraham JB, Toth-Petroczy A, Brock K, Riesselman AJ, Palmedo P, Kang C, Sheridan R, Draizen EJ, Dallago C, Sander C, Marks DS. 2019. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582-1584.

Huang KL, Ma GJ, Zhang ML, Xiong H, Wu H, Zhao CZ, Liu CS, Jia HX, Chen L, Kjorven JO, Li XB, Ren F. 2018. The ARF7 and ARF19 Transcription Factors Positively Regulate PHOSPHATE STARVATION RESPONSE1 in Arabidopsis Roots. *Plant Physiology* **178**, 413-427.

Illergård K, Ardell DH, Elofsson A. 2009. Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics* **77**, 499-508.

Ishizaki K, Chiyoda S, Yamato KT, Kohchi T. 2008. Agrobacterium-mediated transformation of the haploid liverwort Marchantia polymorpha L., an emerging model for plant biology. *Plant & Cell Physiology* **49**, 1084-1091.

Ishizaki K, Nishihama R, Ueda M, Inoue K, Ishida S, Nishimura Y, Shikanai T, Kohchi T. 2015. Development of gateway binary vector series with four different selection markers for the liverwort marchantia polymorpha. *PLoS One* **10**.

Ishizaki K, Nishihama R, Yamato KT, Kohchi T. 2016. Molecular Genetic Tools and Techniques for Marchantia polymorpha Research. *Plant & Cell Physiology* **57**, 262-270.

Ishizaki K, Nonomura M, Kato H, Yamato KT, Kohchi T. 2012. Visualization of auxin-mediated transcriptional activation using a common auxin-responsive reporter system in the liverwort Marchantia polymorpha. *Journal of Plant Research* **125**, 643-651.

Jing H, Korasick DA, Emenecker RJ, Morffy N, Wilkinson EG, Powers SK, Strader LC. 2022. Regulation of AUXIN RESPONSE FACTOR condensation and nucleo-cytoplasmic partitioning. *Nat Commun* **13**, 4015.

Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpaa MJ, Bonke M, Palin K, Talukder S, Hughes TR, Luscombe NM, Ukkonen E, Taipale J. 2010. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* **20**, 861-873.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589.

Kang M, Kim S, Kim HJ, Shrestha P, Yun J-h, Phee B-K, Lee W, Nam HG, Chang I. 2018. The C-Domain of the NAC Transcription Factor ANAC019 Is Necessary for pH-Tuned DNA Binding through a Histidine Switch in the N-Domain. *Cell Reports* **22**, 1141-1150.

Kasajima I, Sasaki K, Tanaka Y, Terakawa T, Ohtsubo N. 2013. Large-scale extraction of pure DNA from mature leaves of Cyclamen persicum Mill. and other recalcitrant plants with alkaline polyvinylpolypyrrolidone (PVPP). *Scientia*

Horticulturae **164**, 65-72.

Kato H, Ishizaki K, Kouno M, Shirakawa M, Bowman JL, Nishihama R, Kohchi T. 2015. Auxin-Mediated Transcriptional System with a Minimal Set of Components Is Critical for Morphogenesis through the Life Cycle in Marchantia polymorpha. *PLoS Genetics* **11**, e1005084.

Kato H, Kouno M, Takeda M, Suzuki H, Ishizaki K, Nishihama R, Kohchi T. 2017. The Roles of the Sole Activator-Type Auxin Response Factor in Pattern Formation of Marchantia polymorpha. *Plant & Cell Physiology* **58**, 1642-1651.

Kato H, Mutte SK, Suzuki H, Crespo I, Das S, Radoeva T, Fontana M, Yoshitake Y, Hainiwa E, van den Berg W, Lindhoud S, Ishizaki K, Hohlbein J, Borst JW, Boer DR, Nishihama R, Kohchi T, Weijers D. 2020. Design principles of a minimal auxin response system. *Nat Plants* **6**, 473-482.

Kato H, Nishihama R, Weijers D, Kohchi T. 2018. Evolution of nuclear auxin signaling: lessons from genetic studies with basal land plants. *Journal of Experimental Botany* **69**, 291-301.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907-915.

Kim J. 1997. Protein-protein interactions among the Aux/IAA proteins. *PNAS*.

Kim Y, Park C, Cha S, Han M, Ryu KS, Suh JY. 2020. Determinants of PB1 Domain Interactions in Auxin Response Factor ARF5 and Repressor IAA17. *Journal of Molecular Biology* 10.1016/j.jmb.2020.04.007.

King GJ, Chanson AH, McCallum EJ, Ohme-Takagi M, Byriel K, Hill JM, Martin JL, Mylne JS. 2013. The Arabidopsis B3 domain protein VERNALIZATION1 (VRN1) is involved in processes essential for development, with structural and mutational studies revealing its DNA-binding surface. *Journal of Biological Chemistry* **288**, 3198-3207.

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. 2015. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* **58**, 610-620.

Korasick DA, Westfall CS, Lee SG, Nanao MH, Dumas R, Hagen G, Guilfoyle TJ, Jez JM, Strader LC. 2014. Molecular basis for AUXIN RESPONSE FACTOR protein interaction and the control of auxin response repression. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 5427-5432.

Kosztin D, Bishop TC, Schulten K. 1997. Binding of the estrogen receptor to DNA. The role of waters. *Biophysical Journal* **73**, 557-570.

Kou X, Zhao X, Wu B, Wang C, Wu C, Yang S, Zhou J, Xue Z. 2022. Auxin Response Factors Are Ubiquitous in Plant Growth and Development, and Involved in Crosstalk between Plant Hormones: A Review. *Applied Sciences* **12**.

Koulechova DA, Tripp KW, Horner G, Marqusee S. 2015. When the Scaffold Cannot Be Ignored: The Role of the Hydrophobic Core in Ligand Binding and Specificity. *Journal of Molecular Biology* **427**, 3316-3326.

Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. 2019. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol* **35**, 357-379.

Kubota A, Ishizaki K, Hosaka M, Kohchi T. 2013. Efficient Agrobacterium-mediated transformation of the liverwort Marchantia polymorpha using regenerating thalli. *Biosci Biotechnol Biochem* **77**, 167-172.

Kuhn A, Roosjen M, Mutte S, Dubey SM, Carrasco PC, Monzer A, Kohchi T, Nishihama R, Fendrych M, Friml J, Sprakel J, Weijers D. 2022. A RAF-like kinase mediates a deeply conserved, ultra-rapid auxin response. 10.1101/2022.11.25.517951.

Lai X, Stigliani A, Lucas J, Hugouvieux V, Parcy F, Zubieta C. 2020. Genome-wide binding of SEPALLATA3 and AGAMOUS complexes determined by sequential DNA-affinity purification sequencing. *Nucleic Acids Research* **48**, 9637-9648.

Lai X, Stigliani A, Vachon G, Carles C, Smaczniak C, Zubieta C, Kaufmann K, Parcy F. 2019. Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants. *Mol Plant* **12**, 743-763.

Lakehal A, Chaabouni S, Cavel E, Le Hir R, Ranjan A, Raneshan Z, Novak O, Pacurar DI, Perrone I, Jobert F, Gutierrez L, Bako L, Bellini C. 2019. A Molecular Framework for the Control of Adventitious Rooting by TIR1/AFB2-Aux/IAA-Dependent Auxin Signaling in Arabidopsis. *Mol Plant* **12**, 1499-1514.

Lambert SA, Yang AWH, Sasse A, Cowley G, Albu M, Caddick MX, Morris QD, Weirauch MT, Hughes TR. 2019. Similarity regression predicts evolution of transcription factor sequence specificity. *Nature Genetics* **51**, 981-989.

Lanctot A, Taylor-Teeples M, Oki EA, Nemhauser JL. 2020. Specificity in Auxin Responses Is Not Explained by the Promoter Preferences of Activator ARFs. *Plant Physiology* **182**, 1533.

Lavrekha VV, Levitsky VG, Tsukanov AV, Bogomolov AG, Grigorovich DA, Omelyanchuk N, Ubogoeva EV, Zemlyanskaya EV, Mironova V. 2022. CisCross: A gene list enrichment analysis to predict upstream regulators in Arabidopsis thaliana. *Frontiers in Plant Science* **13**.

Lavy M, Prigge MJ, Tao S, Shain S, Kuo A, Kirchsteiger K, Estelle M. 2016. Constitutive auxin response in Physcomitrella reveals complex interactions between Aux/IAA and ARF proteins. *Elife* **5**.

Lee J, Chang I, Yu W. 2019. Atomic insights into the effects of pathological mutants through the disruption of hydrophobic core in the prion protein. *Scientific Reports* **9**, 19144.

Lehmann J, Jørgensen ME, Fratz S, Müller HM, Kusch J, Scherzer S, Navarro-Retamal C, Mayer D, Böhm J, Konrad KR, Terpitz U, Dreyer I, Mueller TD, Sauer M, Hedrich R, Geiger D, Maierhofer T. 2021. Acidosis-induced activation

of anion channel SLAH3 in the flooding-related stress response of Arabidopsis. *Current Biology* **31**, 3575-3585. e3579.

Levitsky V, Zemlyanskaya E, Oshchepkov D, Podkolodnaya O, Ignatieva E, Grosse I, Mironova V, Merkulova T. 2019. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. *Nucleic Acids Research* **47**, e139.

Li J, Dai X, Zhao Y. 2006. A role for auxin response factor 19 in auxin and ethylene signaling in Arabidopsis. *Plant Physiology* **140**, 899-908.

Li JF, Bush J, Xiong Y, Li L, McCormack M. 2011. Large-scale protein-protein interaction analysis in Arabidopsis mesophyll protoplasts by split firefly luciferase complementation. *PLoS One* **6**, e27364.

Li SB, Xie ZZ, Hu CG, Zhang JZ. 2016. A Review of Auxin Response Factors (ARFs) in Plants. *Front Plant Sci* **7**, 47.

Liao CY, Smet W, Brunoud G, Yoshida S, Vernoux T, Weijers D. 2015. Reporters for sensitive and quantitative measurement of auxin response. *Nature Methods* **12**, 207-210, 202 p following 210.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930.

Lieberman-Lazarovich M, Yahav C, Israeli A, Efroni I. 2019. Deep Conservation of cis-Element Variants Regulating Plant Hormonal Responses. *Plant Cell* **31**, 2559-2572.

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130.

Liu ZB, Ulmasov T, Shi X, Hagen G, Guilfoyle TJ. 1994. Soybean GH3 promoter contains multiple auxin-inducible elements. *Plant Cell* **6**, 645-657.

Louphrasitthiphol P, Siddaway R, Loffreda A, Pogenberg V, Friedrichsen H, Schepsky A, Zeng Z, Lu M, Strub T, Freter R, Lisle R, Suer E, Thomas B, Schuster-Bockler B, Filippakopoulos P, Middleton M, Lu X, Patton EE, Davidson I, Lambert JP, Wilmanns M, Steingrimsson E, Mazza D, Goding CR. 2020. Tuning Transcription Factor Availability through Acetylation-Mediated Genomic Redistribution. *Molecular Cell* **79**, 472-487 e410.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550.

Luscombe NM, Laskowski RA, Thornton JM. 2001. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research* **29**, 2860-2874.

Marino SM, Gladyshev VN. 2010. Cysteine function governs its conservation and degeneration and restricts its utilization on protein surfaces. *Journal of Molecular Biology* **404**, 902-916.

Martin-Arevalillo R, Thevenon E, Jegu F, Vinos-Poyo T, Vernoux T, Parcy F, Dumas R. 2019. Evolution of the Auxin Response Factors from charophyte ancestors. *PLoS Genetics* **15**, e1008400.

Martin-Pizarro C, Trivino JC, Pose D. 2019. Functional analysis of the TM6 MADS-box gene in the octoploid strawberry by CRISPR/Cas9-directed mutagenesis. *Journal of Experimental Botany* **70**, 885-895.

Mazmanian K, Sargsyan K, Grauffel C, Dudev T, Lim C. 2016. Preferred Hydrogen-Bonding Partners of Cysteine: Implications for Regulating Cys Functions. *The Journal of Physical Chemistry B* **120**, 10288-10296.

McCourt RM, Delwiche CF, Karol KG. 2004. Charophyte algae and land plant origins. *Trends in Ecology & Evolution* **19**, 661-666.

McCourt RM, Lewis LA, Strother PK, Delwiche CF, Wickett NJ, de Vries J, Bowman JL. 2023. Green land: Multiple perspectives on green algal evolution and the earliest land plants. *American Journal of Botany* **110**.

Moi D, Bernard C, Steinegger M, Nevers Y, Langleib M, Dessimoz C. 2023. 10.1101/2023.09.19.558401.

Mueller AL, Corbi-Verge C, Giganti DO, Ichikawa DM, Spencer JM, MacRae M, Garton M, Kim PM, Noyes MB. 2021. The geometric influence on the Cys2His2 zinc finger domain and functional plasticity. *Nucleic Acids Research* **48**, 6382-6402.

Muto H, Nagao I, Demura T, Fukuda H, Kinjo M, Yamamoto KT. 2006. Fluorescence cross-correlation analyses of the molecular interaction between an Aux/IAA protein, MSG2/IAA19, and protein-protein interaction domains of auxin response factors of arabidopsis expressed in HeLa cells. *Plant & Cell Physiology* **47**, 1095-1101.

Mutte SK, Kato H, Rothfels C, Melkonian M, Wong GK, Weijers D. 2018. Origin and evolution of the nuclear auxin response system. *Elife* **7**.

Nagpal P, Ellis CM, Weber H, Ploense SE, Barkawi LS, Guilfoyle TJ, Hagen G, Alonso JM, Cohen JD, Farmer EE, Ecker JR, Reed JW. 2005. Auxin response factors ARF6 and ARF8 promote jasmonic acid production and flower maturation. *Development* **132**, 4107-4118.

Nanao MH, Vinos-Poyo T, Brunoud G, Thevenon E, Mazzoleni M, Mast D, Laine S, Wang S, Hagen G, Li H, Guilfoyle TJ, Parcy F, Vernoux T, Dumas R. 2014. Structural basis for oligomerization of auxin transcriptional regulators. *Nat Commun* **5**, 3617.

Narise T, Kobayashi K, Baba S, Shimojima M, Masuda S, Fukaki H, Ohta H. 2010. Involvement of auxin signaling mediated

by IAA14 and ARF7/19 in membrane lipid remodeling during phosphate starvation. *Plant Molecular Biology* **72**, 533-544.

Nishimasu H, Yamano T, Gao L, Zhang F, Ishitani R, Nureki O. 2017. Structural Basis for the Altered PAM Recognition by Engineered CRISPR-Cpf1. *Molecular Cell* **67**, 139-147.e132.

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EE, Taipale J. 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **4**.

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008a. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**, 1277-1289.

Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA. 2008b. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Research* **36**, 2547-2560.

O'Malley RC, Huang SC, 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.

Okushima Y, Fukaki H, Onoda M, Theologis A, Tasaka M. 2007. ARF7 and ARF19 regulate lateral root formation via direct activation of LBD/ASL genes in Arabidopsis. *Plant Cell* **19**, 118-130.

Okushima Y, Overvoorde PJ, Arima K, Alonso JM, Chan A, Chang C, Ecker JR, Hughes B, Lui A, Nguyen D, Onodera C, Quach H, Smith A, Yu G, Theologis A. 2005. Functional genomic analysis of the AUXIN RESPONSE FACTOR gene family members in Arabidopsis thaliana: unique and overlapping functions of ARF7 and ARF19. *Plant Cell* **17**, 444-463.

One Thousand Plant Transcriptomes I. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679-685.

Osakabe Y, Osakabe K. 2015. Genome Editing with Engineered Nucleases in Plants. *Plant and Cell Physiology* **56**, 389-400.

Ouellet F, Overvoorde PJ, Theologis A. 2001. IAA17/AXR3: biochemical insight into an auxin mutant phenotype. *Plant Cell* **13**, 829-841.

Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, Burley SK. 1999. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes & Development* **13**, 3217-3230.

Perez JC, Fordyce PM, Lohse MB, Hanson-Smith V, DeRisi JL, Johnson AD. 2014. How duplicated transcription regulators can diversify to govern the expression of nonoverlapping sets of genes. *Genes & Development* **28**, 1272-1277.

Persikov AV, Wetzel JL, Rowland EF, Oakes BL, Xu DJ, Singh M, Noyes MB. 2015. A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. *Nucleic Acids Research* **43**, 1965-1984.

Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research* **24**, 2033-2040.

Pierre-Jerome E, Moss BL, Lanctot A, Hageman A, Nemhauser JL. 2016. Functional analysis of molecular interactions in synthetic auxin response circuits. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 11354-11359.

Ponting CP, Ito T, Moscat J, Diaz-Meco MT, Inagaki F, Sumimoto H. 2002. OPR, PC and AID: all in the PB1 family. *Trends Biochem Sci* **27**, 10.

Powers SK, Holehouse AS, Korasick DA, Schreiber KH, Clark NM, Jing H, Emenecker R, Han S, Tycksen E, Hwang I, Sozzani R, Jez JM, Pappu RV, Strader LC. 2019. Nucleo-cytoplasmic Partitioning of ARF Proteins Controls Auxin Responses in Arabidopsis thaliana. *Molecular Cell* **76**, 177-190 e175.

Rademacher EH, Lokerse AS, Schlereth A, Llavata-Peris CI, Bayer M, Kientz M, Freire Rios A, Borst JW, Lukowitz W, Jurgens G, Weijers D. 2012. Different auxin response machineries control distinct cell fates in the early plant embryo. *Developmental Cell* **22**, 211-222.

Rademacher EH, Moller B, Lokerse AS, Llavata-Peris CI, van den Berg W, Weijers D. 2011. A cellular expression map of the Arabidopsis AUXIN RESPONSE FACTOR gene family. *Plant Journal* **68**, 597-606.

Raditsa VV, Tsukanov AV, Bogomolov AG, Levistsky VG. 2023. Genomic background sequences systematically outperform synthetic ones in de novo motif discovery for ChIP-seq data. *Front Genet.* (under review).

Reece-Hoyes JS, Pons C, Diallo A, Mori A, Shrestha S, Kadreppa S, Nelson J, Diprima S, Dricot A, Lajoie BR, Ribeiro PS, Weirauch MT, Hill DE, Hughes TR, Myers CL, Walhout AJ. 2013. Extensive rewiring and complex evolutionary dynamics in a C. elegans multiparameter transcription factor network. *Molecular Cell* **51**, 116-127.

Richter R, Behringer C, Zourelidou M, Schwechheimer C. 2013. Convergence of auxin and gibberellin signaling on the regulation of the GATA transcription factors GNC and GNL in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 13192-13197.

Rieu P, Turchi L, Thevenon E, Zarkadas E, Nanao M, Chahtane H, Tichtinsky G, Lucas J, Blanc-Mathieu R, Zubieta C, Schoehn G, Parcy F. 2023. The F-box protein UFO controls flower development by redirecting the master transcription factor LEAFY to new cis-elements. *Nat Plants* **9**, 315-329.

Roosjen M, Paque S, Weijers D. 2018. Auxin Response Factors: output control in auxin biology. *Journal of Experimental Botany* **69**, 179-188.

Saint-Marcoux D, Proust H, Dolan L, Langdale JA. 2015. Identification of Reference Genes for Real-Time Quantitative PCR Experiments in the Liverwort <italic>Marchantia polymorpha</italic>. *PLoS One* **10**, e0118678.

Salmon J, Ramos J, Callis J. 2008. Degradation of the auxin response factor ARF1. *Plant Journal* **54**, 118-128.

Sanjana NE, Cong L, Zhou Y, Cunniff MM, Feng G, Zhang F. 2012. A transcription activator-like effector toolbox for genome engineering. *Nat Protoc* **7**, 171-192.

Sasnauskas G, Kauneckaitė K, Siksnys V. 2018a. Structural basis of DNA target recognition by the B3 domain of Arabidopsis epigenome reader VAL1. *Nucleic Acids Research* **46**, 4316-4324.

Sasnauskas G, Manakova E, Lapėnas K, Kauneckaitė K, Siksnys V. 2018b. DNA recognition by Arabidopsis transcription factors ABI3 and NGA1. *FEBS Journal* **285**, 4041-4059.

Sayou C, Monniaux M, Nanao M, Moyroud E. 2014. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificty. *Science* **343**, 645-648.

Schmitz JF, Zimmer F, Bornberg-Bauer E. 2016. Mechanisms of transcription factor evolution in Metazoa. *Nucleic Acids Research* **44**, 6287-6297.

Shen C, Wang S, Bai Y, Wu Y, Zhang S, Chen M, Guilfoyle TJ, Wu P, Qi Y. 2010. Functional analysis of the structural domain of ARF proteins in rice (Oryza sativa L.). *Journal of Experimental Botany* **61**, 3971-3981.

Skene PJ, Henikoff S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, e21856.

Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, Mann RS. 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**, 1270-1282.

Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordan R, Rohs R. 2014. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* **39**, 381-399.

Sloan J, Hakenjos JP, Gebert M, Ermakova O, Gumiero A, Stier G, Wild K, Sinning I, Lohmann JU. 2020. Structural basis for the complex DNA binding behavior of the plant stem cell regulator WUSCHEL. *Nat Commun* **11**, 2223.

Smaczniak C, Angenent GC, Kaufmann K. 2017. SELEX-Seq: A Method to Determine DNA Binding Specificities of Plant Transcription Factors. *Methods Mol Biol* **1629**, 67-82.

Starr TN, Picton LK, Thornton JW. 2017. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409-413.

Sterken MG, Nijveen H, van Zanten M, Jiménez-Gómez JM, Geshnizjani N, Willems LAJ, Rienstra J, Hilhorst HWM, Ligterink W, Snoek BL. 2023. Plasticity of maternal environment-dependent expression-QTLs of tomato seeds. *Theoretical and Applied Genetics* **136**, 28.

Stigliani A, Martin-Arevalillo R, Lucas J, Bessy A, Vinos-Poyo T, Mironova V, Vernoux T, Dumas R, Parcy F. 2019. Capturing Auxin Response Factors Syntax Using DNA Binding Models. *Mol Plant* **12**, 822-832.

Sugano SS, Shirakawa M, Takagi J, Matsuda Y, Shimada T, Hara-Nishimura I, Kohchi T. 2014. CRISPR/Cas9-mediated targeted mutagenesis in the liverwort Marchantia polymorpha L. *Plant & Cell Physiology* **55**, 475-481.

Suzuki H, Kato H, Iwano M, Nishihama R, Kohchi T. 2023. Auxin signaling is essential for organogenesis but not for cell survival in the liverwort Marchantia polymorpha. *Plant Cell* **35**, 1058-1075.

Swaminathan K, Peterson K, Jack T. 2008. The plant B3 superfamily. *Trends in Plant Science* **13**, 647-655.

Tabata R, Ikezaki M, Fujibe T, Aida M, Tian CE, Ueno Y, Yamamoto KT, Machida Y, Nakamura K, Ishiguro S. 2010. Arabidopsis auxin response factor6 and 8 regulate jasmonic acid biosynthesis and floral organ development via repression of class 1 KNOX genes. *Plant & Cell Physiology* **51**, 164-175.

Tan X, Calderon-Villalobos LIA, Sharon M, Zheng C, Robinson CV, Estelle M, Zheng N. 2007. Mechanism of auxin perception by the TIR1 ubiquitin ligase. *Nature* **446**, 640.

Taylor SC, Nadeau K, Abbasi M, Lachance C, Nguyen M, Fenrich J. 2019. The Ultimate qPCR Experiment: Producing Publication Quality, Reproducible Data the First Time. *Trends in Biotechnology* 10.1016/j.tibtech.2018.12.002.

Tiwari SB. 2001. AUX/IAA Proteins Are Active Repressors, and Their Stability and Activity Are Modulated by Auxin. *The Plant Cell Online* **13**, 2809-2822.

Tiwari SB, Hagen G, Guilfoyle T. 2003. The roles of auxin response factor domains in auxin-responsive transcription. *Plant Cell* **15**, 533-543.

Tu X, Ren S, Shen W, Li J, Li Y, Li C, Li Y, Zong Z, Xie W, Grierson D, Fei Z, Giovannoni J, Li P, Zhong S. 2022. Limited conservation in cross-species comparison of GLK transcription factor binding suggested wide-spread cistrome divergence. *Nature Communications* **13**.

Tugrul M, Paixao T, Barton NH, Tkacik G. 2015. Dynamics of Transcription Factor Binding Site Evolution. *PLoS Genetics* **11**, e1005639.

Ulmasov T, Hagen G, Guilfoyle TJ. 1997a. ARF1, a transcription factor that binds to auxin response elements. *Science* **276**, 1865-1868.

Ulmasov T, Hagen G, Guilfoyle TJ. 1999a. Activation and repression of transcription by auxin-response factors. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 5844-5849.

Ulmasov T, Hagen G, Guilfoyle TJ. 1999b. Dimerization and DNA binding of auxin response factors. *Plant Journal* **19**, 309-319.
Ulmasov T, Liu ZB, Hagen G, Guilfoyle TJ. 1995. Composite structure of auxin response elements. *Plant Cell* **7**, 1611-1623.
Ulmasov T, Murfett J, Hagen G, Guilfoyle TJ. 1997b. Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *Plant Cell* **9**, 1963-1971.
van Dijk M, Visscher KM, Kastritis PL, Bonvin AM. 2013. Solvated protein-DNA docking using HADDOCK. *Journal of Biomolecular NMR* **56**, 51-63.
van Zundert GCP, Rodrigues J, Trellet M, Schmitz C, Kastritis PL, Karaca E, Melquiond ASJ, van Dijk M, de Vries SJ, Bonvin A. 2016. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology* **428**, 720-725.
Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. 2021. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439-D444.
Ventura S, Cristina Vega M, Lacroix E, Angrand I, Spagnolo L, Serrano L. 2002. Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nature Structural Biology* **9**, 485-493.
Vuzman D, Azia A, Levy Y. 2010a. Searching DNA via a "Monkey Bar" mechanism: the significance of disordered tails. *Journal of Molecular Biology* **396**, 674-684.
Vuzman D, Polonsky M, Levy Y. 2010b. Facilitated DNA search by multidomain transcription factors: cross talk via a flexible linker. *Biophysical Journal* **99**, 1202-1211.
Walcher CL, Nemhauser JL. 2012. Bipartite promoter element required for auxin response. *Plant Physiology* **158**, 273-282.
Wang J, Yan DW, Yuan TT, Gao X, Lu YT. 2013. A gain-of-function mutation in IAA8 alters Arabidopsis floral organ development by change of jasmonic acid level. *Plant Molecular Biology* **82**, 71-83.
Wang L, Wan MC, Liao RY, Xu J, Xu ZG, Xue HC, Mai YX, Wang JW. 2023. The maturation and aging trajectory of Marchantia polymorpha at single-cell resolution. *Developmental Cell* 10.1016/j.devcel.2023.05.014.
Wanninger A. 2024. Hox, homology, and parsimony: An organismal perspective. *Seminars in Cell & Developmental Biology* **152-153**, 16-23.
Wei MT, Chang YC, Shimobayashi SF, Shin Y, Strom AR, Brangwynne CP. 2020. Nucleated transcriptional condensates amplify gene expression. *Nature Cell Biology* **22**, 1187-1196.
Weijers D, Benkova E, Jager KE, Schlereth A, Hamann T, Kientz M, Wilmoth JC, Reed JW, Jurgens G. 2005. Developmental specificity of auxin response by pairs of ARF and Aux/IAA transcriptional regulators. *EMBO Journal* **24**, 1874-1885.
Weijers D, Wagner D. 2016. Transcriptional Responses to the Auxin Hormone. *Annual Review of Plant Biology*, Vol. 67, 539-574.
Wilmoth JC, Wang S, Tiwari SB, Joshi AD, Hagen G, Guilfoyle TJ, Alonso JM, Ecker JR, Reed JW. 2005. NPH4/ARF7 and ARF19 promote leaf expansion and auxin-induced lateral root formation. *Plant Journal* **43**, 118-130.
Wilson KA, Kellie JL, Wetmore SD. 2014. DNA-protein pi-interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Research* **42**, 6726-6741.
Wilson KA, Wells RA, Abendong MN, Anderson CB, Kung RW, Wetmore SD. 2016. Landscape of π-π and sugar-π contacts in DNA-protein interactions. *Journal of Biomolecular Structure and Dynamics* **34**, 184-200.
Wu B, Zhang M, Su S, Liu H, Gan J, Ma J. 2018. Structural insight into the role of VAL1 B3 domain for targeting to FLC locus in Arabidopsis thaliana. *Biochemical and Biophysical Research Communications* **501**, 415-422.
Wu L, Liu H, Xu Y, Nie Y. 2023. Entering an Era of Protein Structuromics. *Biochemistry* **62**, 3167-3169.
Wu MF, Tian Q, Reed JW. 2006. Arabidopsis microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. *Development* **133**, 4211-4218.
Xu M, Tang D, Cheng X, Zhang J, Tang Y, Tao Q, Shi W, You A, Gu M, Cheng Z, Yu H. 2019. OsPINOID Regulates Stigma and Ovule Initiation through Maintenance of the Floral Meristem by Auxin Signaling. *Plant Physiology* **180**, 952-965.
Xu N, Hagen G, Guilfoyle T. 1997. Multiple auxin response modules in the soybean SAUR 15A promoter. *Plant Science* **126**, 193-201.
Yamaguchi N, Wu M-F, Winter Cara M, Berns Markus C, Nole-Wilson S, Yamaguchi A, Coupland G, Krizek Beth A, Wagner D. 2013. A Molecular Framework for Auxin-Mediated Initiation of Flower Primordia. *Developmental Cell* **24**, 271-282.
Yamasaki K, Kigawa T, Inoue M, Tateno M, Yamasaki T, Yabuki T, Aoki M, Seki E, Matsuda T, Tomo Y, Hayami N, Terada T, Shirouzu M, Osanai T, Tanaka A, Seki M, Shinozaki K, Yokoyama S. 2004. Solution Structure of the B3 DNA Binding Domain of the Arabidopsis Cold-Responsive Transcription Factor RAV1. *The Plant Cell* **16**, 3448-3459.
Yang C, Liu X, Li D, Zhu X, Wei Z, Feng Z, Zhang L, He J, Mou C, Jiang L, Wan J. 2019. OsLUGL is involved in the regulating auxin level and OsARFs expression in rice (Oryza sativa L.). *Plant Science* **288**, 110239.
Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, Nitta KR, Taipale M, Popov A, Ginno PA, Domcke S, Yan J, Schubeler D, Vinson C, Taipale J. 2017. Impact of cytosine methylation

on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239.

Yu Y, Tang W, Lin W, Li W, Zhou X, Li Y, Chen R, Zheng R, Qin G, Cao W, Perez-Henriquez P, Huang R, Ma J, Qiu Q, Xu Z, Zou A, Lin J, Jiang L, Xu T, Yang Z. 2023. ABLs and TMKs are co-receptors for extracellular auxin. *Cell* **186**, 5457-5471 e5417.

Zeiske T, Baburajendran N, Kaczynska A, Brasch J, Palmer AG, 3rd, Shapiro L, Honig B, Mann RS. 2018. Intrinsic DNA Shape Accounts for Affinity Differences between Hox-Cofactor Binding Sites. *Cell Rep* **24**, 2221-2230.

Zemlyanskaya EV, Wiebe DS, Omelyanchuk NA, Levitsky VG, Mironova VV. 2016. Meta-analysis of transcriptome data identified TGTCNN motif variants associated with the response to plant hormone auxin in Arabidopsis thaliana L. *J Bioinform Comput Biol* **14**, 1641009.

Zhang Z, Coenen H, Ruelens P, Hazarika RR, Al Hindi T, Oguis GK, Vandeperre A, van Noort V, Geuten K. 2018. Resurrected Protein Interaction Networks Reveal the Innovation Potential of Ancient Whole-Genome Duplication. *The Plant Cell* **30**, 2741-2760.

Summary

Auxin is a crucial hormone involved in virtually all growth and development processes in land plants. Key to the cellular auxin signalling transduction pathway are the Auxin Response Factor (ARF) transcription factors (TFs), which determine the set of genes that are auxin-responsive by their sequence-specific DNA-binding to auxin response elements (AuxRE). This thesis presents a comprehensive analysis of ARF AuxRE binding preferences across the plant kingdom.

Chapter 1 reports on the biochemical mechanisms that facilitate TF DNA-binding and the evolutionary dynamics that drive their DNA-binding specificities. We introduce the nuclear auxin signalling pathway (NAP) and the critical role of AuxRE binding by ARFs. We establish the evolutionary trajectory of the ARFs within the streptophytes (the land plants and their closer relatives, the streptophyte algae), introduce the NAP of the liverwort *Marchantia polymorpha*, and establish the scope of the thesis.

Chapter 2 reviews the DNA-binding specificities of the ARFs. We review the structure of the ARF DNA-binding domain (DBD) and how two subdomains specify DNA-specificity: the B3 domain that physically interacts with the AuxRE, and the dimerization domain (DD) that defines the necessary and tolerated spacing between two AuxRE. We explore the divergent spacing preferences among the ARF families – A, B and C-ARFs – and propose future research suggestions.

The genomic revolution led to hundreds of sequenced plant genomes and transcriptomes, covering over a billion years of plant evolution. In parallel, advances in artificial intelligence (AI) led to the breakthrough development of AI systems that can predict high-quality protein structures for virtually any sequence. In **Chapter 3**, we capitalize on these developments and examine evolutionary conservation at the level of individual residues and the ARF DNA-binding domain (DBD) as a whole. We reveal that the residues of the B3 domain, the DBD subdomain that physically interacts with the AuxRE, have been strongly conserved during evolution within the plant kingdom, and that several DNA-binding residues are evolutionarily coupled with others. We find that the overall DBD structure is conserved across all three ARF families, likely due to the conservation of structural, buried residues. Lastly, we observe substantial variation in the residues involved in homodimerization and in B3 flexibility (the B3-hinge).

To understand the degree of evolutionary conservation in AuxRE specificities among the A-class ARFs, we characterize the DNA-binding preferences of the single A-ARF from *M. polymorpha,* MpARF1, in **Chapter 4**. DNA Affinity Purification (DAP-seq) reveals that the MpARF1-B3 binding preferably binds TGTCGG AuxRE, a conserved preference shared with angiosperm A-ARFs. A combination of DAP-seq and motif enrichment analysis in promoters of auxin-responsive genes in *M. polymorpha* reveals that the

spacing preferences of MpARF1 are different from those of angiosperm ARFs. Structural modelling supports the possibility that the newly identified DR2 motif is bound by a cooperative MpARF1 homodimer.

The strong conservation of the DNA-binding residues and the preference of MpARF1 for TGTCGG AuxRE suggested the ARF-AuxRE interface to be highly evolutionary conserved. In **Chapter 5**, we investigate the evolvability of this interface using *in vivo* complementation assays with MpARF1. We find that mutations in the residues interacting with the TGTC-core generally render MpARF1 nonfunctional, with one notable exception. Surprisingly, mutations in the residue that binds G5 and G6 (of TGTCGG) are tolerated, as tested in a series of phenotyping experiments. Two mutants with delayed growth are further investigated using a transcriptomics approach and are found to have reduced sensitivity to exogenous auxin. The setup of the experiment also allows us to confirm that MpARF1 is the sole TF in *M. polymorpha* responsible for transducing the auxin signal. Molecular docking suggests mechanisms underlying how these mutations affect DNA-binding.

A previous study indicated that the DBD's of the A-, B- and C-ARF of *M. polymorpha* are functionally distinct. However, that study could not distinguish between AuxRE- and spacing-preferences of the ARFs. The preceding chapters reveal the strong conservation and low evolvability of the B3-AuxRE interface. We therefore uncouple AuxRE and spacing preferences in a series of B3 domain swaps with MpARF1 in **Chapter 6**. The B3 domains of the A-ARFs from *Arabidopsis thaliana*, the B-ARF from *M. polymorpha* and the A/B-ARFs from streptophyte algae are all functionally similar to the MpARF1 B3 domain, thereby indicating that the AuxRE preferences of these ARFs are well conserved across multiple streptophyte lineages. Furthermore, it indicates that the DNA-binding preferences of the A/B-ARFs were acquired by their last common ancestor and that they have been retained in the expanded A-ARF family of *A. thaliana*. In contrast, B3 domains of C-ARFs only partially maintain the capacity to replace the MpARF1 B3 domain, indicating that there is a functional divergence from the A- and B-ARFs in their DNA-binding preferences.

Chapter 7 synthesizes and discusses the findings within the broader perspective of the evolution of the ARF family. The results of this thesis collectively show how conserved the ARF-AuxRE interface is, particularly for the A/B-ARFs. We also highlighted future research directions and open questions.

Acknowledgments

Ever since high school, when I read about insulin production in *E.coli*, I knew I wanted to pursue a scientific career in molecular biology. Many years later, with a small detour as a technician, I am very close to the conclusion of this goal. Ready to finally end one chapter and begin the next. I thought a PhD project was only a training in scientific thinking, but it turned out to be so much more than that and I needed to develop my project-management, collaboration and supervision skills, and more than anything, develop mental resilience. The saying goes that it takes a village to raise a child. It seems it equally takes an entire group to train a PhD. I would never have managed if I did not get help along the way from many colleagues and friends that I want to thank here.

First and foremost I want to thank my supervisor **Dolf Weijers**. At the coffee break from your seminar during my first MSc course I asked you whether we could write a proposal together, to which you said yes without hesitation. That concept turned into a successful proposal and now a great ending to an exciting project. Your openness to give chances to people, your enthusiasm when thinking of ambitious topics, and your first priority being the development of the person, truly sets you apart in my eyes. You have supported me scientifically and personally and I am grateful and proud that you were my mentor.

I am also extremely grateful to my second supervisor **Jorge Hernández García**. You became my second supervisor at a very tough time in my PhD, and you managed reignite my motivation by your support with experiments, your awesome ideas and your positive (albeit slightly sarcastic) motivation. Chapter 6 was only possible because of your teamwork, and every chapter of my thesis is better because of your input. It was also a ton of fun to come up with the craziest, most outrageous experiments we could think of, or just talk about random stuff.

I cannot thank my paranimphs enough for their personal and scientific support. Lord Duke **Sjoerd Woudenberg** of Woudenberg, it has been a huge pleasure to share the office with you. You always had an answer to any of my random plant questions, which enriched my own understanding and appreciation of evolution. This has been particularly handy as I dove into the, for me unfamiliar, world of early divergent land plants (just a joke! Please remain my paranimph!). On a personal note, I will really miss joking around, your sharp and insightful scientific suggestions, and your selflessness when helping people. **Sumanth Mutte**, I would have needed twice the time for my project without your bioinformatics wizardry and keen insights. You helped me develop my scripting skills (although sometimes at a cost to your own mental sanity, i.e., the *run alpafold* debacle), and my mental resilience at the squash court. I hope we will have many more beer-induced long discussions about food, culture, business and science in

the future too. Both of you, thanks for being such great friends.

Of course Sjoerd and I were not alone in the office. **Adrie Westphal**, **Annemarie Bok**, **Bel Koopal**, and **Pilar Bobadilla**, we definitely had the best and loudest office. I really treasure the moments when normal conversations went completely off the rails. Adrie, I had zero structural biology experience in the beginning and overloaded you with questions. Thanks for answering all of them, simultaneously teaching me about astrophysics, and for all of your jokes. To our old office mate **Dong Hao**, thanks for all the fun times in the office and on the squash and badminton courts. I will never forget the cola-marinated chicken feet, and likewise hope you still make dishes with Italian noodles.

Doing a PhD is hard, but sharing struggles, ideas and support with the other members of BIC made it a lot easier. **Ping Harvanichvech**, we have spend many fantastic moments, trips, coffees, restaurant visits, desserts, and croissants together, and you have always given me great advice. Thanks for being such a good and supportive friend, and I cannot wait to visit Thailand. Please do not praise my Risotto anymore. **Aaron Ang**, the moment you are finally acknowledged in Dolfs selfies is sadly when I finish my PhD. It is sad because you bring great energy and humor to a group, which also made our adventures to Leiden super fun. Maybe one day I will learn which animal I am? **Martijn de Roij**, your microscopy images were truly stunning! Thanks for always being up for some jokes and fun. **Polet Carrillo-Carrasco**, thanks for all the chats, support and just for always bringing positivity with you. **Andriy Volkov**, sneaking in a gouda to a whisky event and accidentally getting hammered was just the thing I needed! Thanks for always being willing to help others. **Joris Sprakel**, thanks for reaching out a helping hand to someone in a tough spot. **André Kuhn**, **Cecilia Borassi**, **João Ramalho**, and **Victor Jones**, thanks for always being open for a chat and sharing your postdoc wisdom and advice. **Hirotaka Kato**, **Keita Tanaka** and **Kuan-Ju Lu**, thanks for all your help in the beginning of my project and for introducing me to the wonderful world of liverworts and biochemistry. Tanaka-san in particular, with your skills and keen scientific insights it was an honor to learn from you and share a project together. **Sergio Martin Ramirez**, **Feras Hadid**, **Mattia Fontana**, **Shubhajit Das**, **Daan Swarts** and all other wonderful members of the group, thank you all for making my time at BIC so awesome.

I also want to thank all the support staff of our lab, without whom the lab would turn into anarchy. My thanks to **Iraes Rabbers** and **Lisette Nitsch** to find a replacement on such short notice, **Willy van der Berg** for always fixing the devices and being a endless source of knowledge, **Che-Yang Liao** for his open mind and always being up for a beer, **Cathy Albrecht** for sharing lab-hacks to work smarter instead of more, and **Mark Roosjen** and **Simon Lindhoud** for all their humor. You guys flanking my lab bench was truly a one-of-a-kind experience. You can finally stop counting the days till I leave ;). Simon, it was also a lot of fun to skate together and have lunch with your family. Perhaps one day

I might even miss your sense of humor ;). Finally, a huge thanks to **Laura van Egmond** because without you the lab would stop functioning. Thanks for all the help booking (and sometimes rebooking) things and for going above and beyond to help all of us.

One of the fun aspects of the PhD was to mentor and supervise students. **Erik van der Velde**, **Yiyun Li** and **Simone van Roosmalen**,I hope you learned as much from me as I did from you and I am happy to see all your contributions in chapter 5. It was really fun to supervise you and see your development. Erik, on your first day we went into Covid-19 lockdown and had to think of an alternative project. You used your Python wizardry on the HADDOCK structures that became critical in understanding the mutant phenotypes. Yiyun, writing an MSc thesis can be really stressful, especially while in a 3-week Covid isolation in a hotel in China. Still, we had a ton of fun together and I hope we can still do a hotpot at some point? ;) Simone, your most crucial B1H work, where you showed almost no residue can replace R194, P197 or R199 for TGTC-binding, unfortunately did not make it into the thesis because of a stupid yeast contamination (although now that I wrote it here its technically still in the thesis :D). Long after you left, people still referenced your jokes and your kind advice to people with colds or flu, which is a testament to how much fun it was to have you in our group.

DAP-seq has been challenging me since day one, until I finally managed with the help of collaborators in my last year. **Pingtao Ding** and **Yuli Ding**, thanks for hosting me and Aaron in your lab for a few weeks and being so generous with your time. **Victor Levitsky** and **Victoria Mironova**, thanks a lot for your help with the data analysis. You extracted so much information from what we thought to be a mediocre dataset. Without any of your help, there would not have been a chapter 4. Shootout to **Merijn Kerstins** and **Kai Thoris** for your suggestions for DAP-seq.

With rampant misinformation eroding trust in science, open and honest science communication is all the more important. I therefore joined the GeneSprout Initiative (GS) with **Damian Boer**, **Ellen Slaman**, **William de Martinez**, **Nikita Sajeev** and **Vera Veltkamp**. It was such an amazing experience to work with you all. William, also many thanks for the excellent hamburgers. Nikita, for almost one year the Dutch GS was just the two of us. Now that I see the achievements of the current members I could not be happier that we persevered. I also enjoyed our conversations (and get your advice) on science, careers and personal topics. **Savio Rodriguez**, **Elena del Pup** and **Halford Dace**, it has been an absolute pleasure to work with you and to get inspired by your ideas and ways of thinking. Savio, it was awesome to finally find another MMA fan. Also my big thanks to the Belgian sproutjes **Nick Vangheluwe** and **Gwen Swinnen**.

My science journey did not start in Biochemistry, but at Plant Physiology. **Jasper Lammers**, somehow we worked at the exact same departments but never at the same time. Thanks for always being up for a cup of coffee, sharing images of random liverworts

and for cheering me up during the writing with your self-deprecating humor. **Gonda Buijs**, I am surprised I had enough brain cells left to do a PhD after all of our whisky and beer tasting events! And like you said, "whatever happens in Wanroij, stays in Wanroij!" Nobody else has to see those pictures xD. That also counts for you **Leo Willems**! Thanks for always arranging access to Radix for me, especially the times you did not know I was coming! **Mariana Silva Artur**, Mariana!, thanks for your advice on my PhD proposal, the career advice, support and for all our chats about life! **Emilie van Dijk Fradin** and **Bas Dekkers**, thank you both for being incredible role models on how to efficiently setup experiments, ask relevant questions and generally tons of practical advice, all of which helped me tremendously during my PhD. My special thanks go out to **Elise Serin**, **Nafiseh Geshnizjani**, **Harm Nijveen**, **Bing Bai**, **Eva van Zelm**, **Melissa Leeggangers**, **Shuang Song**, **Henk Hilhorst**, and **Leónie Bentsink** for creating an environment where I could develop as a scientist and person even before my PhD. Also a thanks to all the members of the old PPH that I could not mention for brevity, it was an amazing experience.

There are three people that deserve a special mention. First, **Wilco Ligterink**, because you are very special! You bring a unique, irreplaceable positive energy to any group, but more than that, my PhD would not even have happened without you. When I send you a panicky email asking for a letter of recommendation for the Graduate Program at 9 pm of the first Sunday after New Years 2018, you send me one that same night so I did not miss the application deadline of Monday 9 am. Thanks so much for that, and for all your advice and chats over the years. **Richard Immink**, it is rare for a PhD candidate to be able to boast of cleaning dishes at a Sinterklaas party with one of their opponents. It may also be rare for a PhD to thank their opponent in the thesis, but over the last 11 years you have been there whenever I needed advice or support; first at PPH, then during my MSc, and finally as my external supervisor. Having you as my first opponent seems like a proper conclusion of the last decade. Thanks for all your support. Likewise, I want to thank **John van 't Klooster**. You convinced me how cool plants are at the HAN, suggested Wilco for my BSc thesis, helped me with applying to the Graduate Program during my MSc and finally suggested Dolf as my PhD supervisor. You have given me the guidance and direction at every step of my career, for which I am very grateful. Which brings me to the next step: can you suggest me a lab for a PostDoc?

Een PhD project vraagt ontzettend veel tijd en energie, maar gelukkig kom ik tot rust wanneer ik met mijn vrienden afspreek. Zeker met mijn oudste vriendengroep met **Christian de Vries**, **Nick Lindeloo de Jong**, **Nils Rotgans** en **Rik Visser**, iedere keer dat we elkaar zien is het weer als vanouds gezellig en creëren we weer nieuwe verhalen. Tegelijkertijd zorgen jullie ervoor dat ik niet naast mijn schoenen ga lopen met mijn doctorstitel. Nils, super bedankt voor al je design tips bij het maken van dit boekje, het ziet er echt fantastisch uit. Misschien dat ik ooit nog een speciale oplage print waarin *alles* in Comic Sans staat als bedankje. **Jenna Vasington**, thank you so much for the

funny-named-birds coffee mug, it always made me smile while writing. De vrienden van de studies heb ik tijdens de PhD (en Covid) jammer genoeg minder vaak gezien dan ik zou willen, maar **Daan Schreurs**, **Manon Theunissen**, **Nick Pelzers**, **Rosanne van Velthoven**, **Bart Klemann**, **Marc Hendriks**, en **Rick Hoogveld**, het is altijd gezellig met jullie!

Fred, **Lidwien** en **Karsten Rienstra**, zonder jullie had ik dit allemaal nooit kunnen bereiken. Pap en Mam, dankzij jullie onvoorwaardelijke steun en liefde heb ik mijn passie voor biologie altijd kunnen volgen, en daar ben ik ontzettend dankbaar voor. Karsten, niemand geeft me zoveel stof tot nadenken als jij wanneer we in de auto zitten en het over de wereld, politiek en het leven hebben. Die nieuwe inzichten laten zien dat ik nog altijd dingen bij kan leren. Dat komt natuurlijk door die ene punt hoger op jouw Cito toets. Ik ben super trots dat ik jullie zoon en broer ben. **Gül Hatinoğlu**, I am so happy we found each other during our MSc theses at Biochemistry. Since then, your endless support has been the bright light during the tougher moments, and I am super thankful that we shared the PhD life together. Thanks for your feedback on each chapter, and thanks for always making me laugh. *Seni seviyorum*.

Curriculum Vitae



Juriaan Rienstra was born on the 29th of July in Rheden, the Netherlands. In 2009, he enrolled for a BSc in Applied Sciences at the Hogeschool of Arnhem and Nijmegen (HAN). After two years of working with bacteria, yeast and human cells, he pivotted to a major in Plant Biotechnology, graduating in 2013. From 2013 to 2017, he worked as a technician in the department of Plant Physiology at Wageningen University & Research (WUR), under the supervision of Wilco Ligterink and Emilie van Dijk-Fradin. In these projects he worked with genetic populations of *Arabidopsis thaliana* and tomato to (fine-) map genes responsible for seed germination and abiotic stress tolerance.

In 2017, knowing he wanted to get his PhD degree, he enrolled for the MSc in Plant Biotechnology from the WUR. That was where Juriaan met Prof. Dr. Dolf Weijers and together they succesfully wrote and secured a PhD project proposal titled: *Laying the foundation of plant shape by design: cracking the code of ARF5 DNA-binding*. Although the project initially aimed to test a hundred engineered ARFs for DNA-binding and altered phenotypes, the project redirected towards unravelling the conservation of DNA-binding preferences during evolution of the ARFs. The results of this project are reported in this thesis. Driven by a passion for molecular biology and structural biology, Juriaan aims to pursue an academic career to further explore and discover the fascinating world of plant evolution.

Publications

Rienstra J, Hernández García J, Weijers D. 2023. To bind or not to bind: how Auxin Response Factors select their target genes. *Journal of Experimental Botany* 74 (22). DOI:10.1093/jxb/erad259.

Snoek BL, Serin E, Nijveen H, Willems LAJ, **Rienstra J**, van Zanten M, Hilhorst HWM, Ligterink W. 2023. Modulation of quantitative trait loci for Arabidopsis thaliana seed performance by the maternal and germination environment. *bioRxiv*. DOI: 10.1101/2023.02.22.529582.

Sterken M, Nijveen H, van Zanten M, Jiménez-Gómez JM, Geshnizjani N, Willems LAJ, **Rienstra J**, Hilhorst HWM, Ligterink W, Snoek BL. 2023. Plasticity of maternal environment-dependent expression-QTLs of tomato seeds. *Theoretical and Applied Genetics* 138, 28. DOI: 10.1007/s00122-023-04322-0.

Vangheluwe N, Swinnen G, de Koning R, Meyer P, Houben M, Huybrechts M, Sajeev N, **Rienstra J**, Boer D. 2020. Give CRISPR a Chance: the GeneSprout Initiative. *Trends in Plant Science* 25(7). DOI: 10.1016/j.tplants.2020.04.11.

Geschnizjani N, Snoek BL, Willems LAJ, **Rienstra J**, Nijveen H, Hilhorst HWM, Ligterink W. 2020. Detection of QTLs for genotype x environment interactions in tomato seeds and seedlings. *Plant, Cell & Environment* 43(8). DOI: 10.1111/pce.13788.

Silva Artur M, **Rienstra J**, Dennis T, Farrant JM, Ligterink W, Hilhorst HWM. 2019. Structural plasticity of intrinsically disordered LEA proteins from Xerophyta schlechteri provides protection in vitro and in vivo. *Frontiers in Plant Science* 10. DOI: 10.3389/fpls.2019.01272.

Plantenga F, Heuvelink E, **Rienstra J**, Visser RGF, Bachem CWB, Marcelis LFM. 2018. Coincidence of potato CONSTANS (StCol1) expression and light cannot explain night-break repression of tuberization. *Physiologia Plantarum* 167(2). DOI: 10.1111/ppl.12885.

Mohemed N, Charnikhova T, Fradin E, **Rienstra J**, Babiker AGT, Bouwmeester HJ. 2018. Genetic variation in Sorghum bicolor strigolactones and their role in resistance against Striga hermonthica. *Journal of Experimental Botany* 69(9). DOI: 10.1093/jxb/ery041.

Bac-Molenaar JAB, Fradin E, **Rienstra J**, Vreugdenhil D, Keurentjes JJB. 2015. GWA mapping of anthocyanin accumulation reveals balancing selection of MYB90 in Arabidopsis thaliana. *PLoS ONE* 10(11), e0143212. DOI: 10.1371/journal.pone.0143212.

Bac-Molenaar JAB, Fradin E, Becker F, **Rienstra J**, van der Schoot J, Vreugdenhil D, Keurentjes JJB. 2015. Genome-wide association mapping of fertility reduction upon heat stress reveals developmental stage-specific QTLs in Arabidopsis thaliana. *The Plant Cell* 27(7). DOI: 10.1105/tpc.15.00248.

Education Statement of the Graduate School
Experimental Plant Sciences

Issued to: Juriaan Rienstra
Date: 30 April 2024
Group: Biochemistry
University: Wageningen University & Research

1) Start-Up Phase	date	cp
► **First presentation of your project**		
Laying out the foundation of plant shape by design: Cracking the code of ARF5 DNA-binding	01 Oct 2019	1.5
► **Writing or rewriting a project proposal**		
► **Writing a review or book chapter**		
Juriaan Rienstra, Jorge Hernández-García, Dolf Weijers, To bind or not to bind: how AUXIN RESPONSE FACTORs select their target genes, *Journal of Experimental Botany* , Volume 74, Issue 22, 1 December 2023, Pages 6922–6932, https://doi.org/10.1093/jxb/erad259	05 Jul 2023	6.0
► **MSc courses**		
Subtotal Start-Up Phase		7.5

2) Scientific Exposure	date	cp
► **EPS PhD days**		
EPS PhD Get 2 Gether	10-11 Feb 2020	0.6
EPS PhD Get 2 Gether - Online	01-02 Feb 2021	0.4
EPS PhD Get 2 Gether	03-04 May 2022	0.6
► **EPS theme symposia**		
EPS Theme 1 - Developmental Biology of Plants - Online	28 Jan 2021	0.2
EPS Theme 1 - Developmental Biology of Plants	14 Jun 2022	0.3
EPS Theme 1 - Developmental Biology of Plants	03 Oct 2023	0.3
► **Lunteren Days and other national platforms**		
Annual meeting Experimental Plant Sciences - Online	12-13 Apr 2021	0.5
Annual meeting Experimental Plant Sciences	11-12 Apr 2022	0.6
Annual meeting Experimental Plant Sciences	17-18 Apr 2023	0.6
Project Consortium: Kickoff meeting	17 Sep 2019	0.1
Project Consortium: year 1 update	27 Jan 2021	0.1
Project Consortium: year 4 update	21 Feb 2023	0.1
National Open Science Festival	11 Feb 2021	0.3
NWO Life 2021 - Networks in Life	26-28 May 2021	0.6
► **Seminars (series), workshops and symposia**		
Seminars		
Live-cell imaging of the polarization dynamics of the Arabidopsis zygote - Dr. Minako Ueda	4 Sep 2019	0.1
A likely ancestral mechanism for the coordination of transcriptional programs - Prof. Miguel Blazquez	8 Oct 2019	0.1
The genetic basis of body plan evolution in plants - Dr. Jill Harrison	8 Oct 2019	0.1
Structure and function of chromatin in bacteria and Archaea - Prof. Remus Dame	17 Oct 2019	0.1
From transcription to shoot morphogenesis - Prof. Jan Lohman	16 Jan 2020	0.1
Resurrected proteins on the Origin of flowers, the impact of whole genome duplications and origin of hormone signalling - Prof. Koen Geuten	6 Feb 2020	0.1
Calcium signaling at the microtubule cytoskeleton and its role in plant growth regulation - Dr. Katharine Bürstenbinder	19 Feb 2020	0.1
Lateral root formation: from waxes to the cytoskeleton - Prof. Joop Vermeer	19 Feb 2020	0.1
Cell polarity across multicellular kingdoms - Prof. Dolf Weijers	27 May 2020	0.1
Bringing agricultural weeds into the molecular lab - Prof. Dana MacGregor	27 May 2020	0.1
Hormonal regulation of self-organizing and adaptive plant development - Prof. Jiří Friml	24 Jun 2020	0.1
Reconstructing regulation: can biological complexity be reverse engineered? - Nicola Patron	24 Jun 2020	0.1
Vernalization: non-coding transcription and epigenetic switching at FLC - Prof. Caroline Dean	01 Jul 2020	0.1
Shade-avoidance signalling warms up - Prof. Jorge Casal	01 Jul 2020	0.1
Plantae presents: Julia Bailey-Serres and Rashmi Sasidharan	28 Apr 2021	0.1
EMBL Careers webinar | Postdoc positions in Industry	14 May 2021	0.1
MAK'IT Fellows Event: New Breeding Techniques: present and future challenges for sustainable food systems. An open and multi-stakeholder dialogue with new generations	21 May 2021	0.1
EU40: New Breeding Techniques (NBTs): Challenges and Opportunities	09 Jun 2021	0.1
PlantEd webinar: Understanding the potential contributions of genome editing to sustainable agriculture	10 Jun 2021	0.1
Sexual reproduction and sex differentiation in gametophythe generation - Prof. Takayuki Kochi	15 Jun 2021	0.1
Plantae Community: Impact of COVID Pandemic on graduate students mental health	1 Nov 2021	0.1
Nature Research Webcast: How to plan your career as a scientist	4 Nov 2021	0.1
Plantae Community: The science of policy: a guide to political activism for early career plant scientists	15 Nov 2021	0.1
European Commission: New genomic techniques - way forward for safe and sustainable innovation in the agri-food sector	29 Nov 2021	0.1
WEES seminar: Short history of carpel development and evolution - Prof. Annette Becker	16 Dec 2021	0.1
Bayesian Data Analysis: a new perspective in the VLAG domain	7 Mar 2022	0.1
Fundamentals of Knowledge Utilization - CVC	4 Apr 2022	0.1
Exploring intercellular protein transport mechanisms in Nicotiana benthamiana and Marchantia polymorpha - Kuan-Ju Lu	16 May 2022	0.1
Another brick in the wall: regulating cell wall synthesis in plants - Joe Kieber	30 May 2022	0.1
EPS flying seminar: Simon Clark	15 Jun 2022	0.1
Cellular dynamics and transcriptional regulations to establish the body axis in plant embryo - Prof. Yusuke Kimata	17 Jun 2022	0.1
Elsevier webinar: The evolving RNA-seq experiment: advances in data quality, analysis and complexity	16 Aug 2022	0.1
EPSO 15th Plant Science Seminar - Data science for understanding plants	15 Sep 2022	0.2
WEES seminar: Evolution of plant architecture through changes in auxin movement control - Yoan Coudert	24 Nov 2022	0.1
EPS flying seminar: Anne Osborn	02 Dec 2022	0.1
The revolution and evolution of activity-based protein profiling - Nature Research	27 Mar 2023	0.2
The liverwort body plan: from transcription factors to stem cells - Facundo Romani	12 Apr 2023	0.1
Evolution of Receptor Kinases signalling in land plants - Isabel Monte	02 May 2023	0.1
Uncovering how plant roots use hormones as proxies to sense and respond to soil stresses - Prof. Malcolm Bennett	02 May 2023	0.1
Cellular roadmaps: control of cell architecture by dynamic microtubules - Prof. Anna Akhmanova	11 May 2023	0.1
Extracellular auxin: perception and coordination with intracellular auxin - Prof. Zhenbiao Yang	22 Jun 2023	0.1
Learning the grammar of crop regulatory DNA for precise engineering of gene expression - Prof. Christine Queitsch	06 Jul 2023	0.1
Workshops		
Plants and Patents	21 Oct 2019	0.2
Working with model bryophytes - Yoan Coudert	24 Nov 2022	0.1
Symposia		
Gene edited crops; global perspectives and regulation - COGEM	10 Oct 2019	0.3

CONTINUED ON NEXT PAGE

	Symposia		
	Gene edited crops; global perspectives and regulation - COGEM	10 Oct 2019	0.3
	Plant Growth and Development Research Unit meeting	29 Oct 2021	0.2
	Plant Development & Adaptation Research Unit meeting	23 Sep 2022	0.2
	Plant Development & Adaptation Research Unit meeting	10 Feb 2023	0.2
	Plant Development & Adaptation Research Unit meeting	01 Jun 2023	0.2
	Explore the Extremes - Alchimica - Wageningen	09 Jun 2022	0.1
►	**Seminar plus**		
►	**International symposia and congresses**		
	CRISPRcon 2019 (Wageningen, the Netherlands)	20-21 Jun 2019	0.6
	Student & Young-Professional Day - European Biotech Week (Brussels, Belgium)	23 Sep 2019	0.2
	Auxin 2022 (Dubrovnik, Croatia)	02-07 Oct 2022	1.3
	Auxins and Cytokinins in Plant Development International Symposium 2023 (Prague, Czech Republic)	25-29 Jun 2023	1.5
►	**Presentations**		
	Posters		2.0
	Annual meeting Experimental Plant Sciences 2021	12-13 Apr 2021	
	EPS Advanced Course: Transcription Factors and Transcriptional Regulation	13-15 Dec 2021	
	Auxin 2022 (Dubrovnik, Croatia)	02-07 Oct 2022	
	Annual meeting Experimental Plant Sciences 2023	17-18 Apr 2023	
	Auxins and Cytokinins in Plant Development 2023 (Prague, Czech Republic)	25-29 Jun 2023	
	Presentations		1.0
	Understanding and engineering of ARF-DNA interactions in auxin response - Annual meeting Experimental Plant Sciences 2022	11 Apr 2022	
►	**IAB interview**		
►	**Excursions**		
	EPS PhD company visit to Enza Zaden & Seed Meets Technology	29 Sep 2022	0.3
	Subtotal Scientific Exposure		18.1

3) In-Depth Studies		*date*	*cp*
►	**Advanced scientific courses & workshops**		
	EMBL Course: computing skills for reproducible research: software carpentry	19-23 Oct 2020	1.5
	EPS Bioinformatic Introduction Course	05-09 Jul 2021	1.5
	EPS CRISRP/Cas Workshop	13-14 Sep 2021	0.6
	EPS Advanced Course: Trancription Factors and Transcriptional Regulation	13-15 Dec 2021	1.0
	EMBL-EBI Course: Introduction to RNA-seq and functional interpretation	21-25 Feb 2022	1.5
	EPS/VLAG Advanced Course: Microscropy and Spectroscopy in Food and Plant Sciences	05-07 Jul 2022	0.9
	Workshop in Streptophyte Algae	15-16 May 2023	0.5
	EMBO Workshop: An integrated view of early land plant evolution	8-10 Nov 2023	0.6
►	**Journal club**		
	Held biweekly in Dolf Weijers lab	2019-2024	3.0
►	**Individual research training**		
	Training in DAP-seq in the lab of Dr. Pingtao Ding (Leiden University)	04-15 Jul 2022	3.0
	Subtotal In-Depth Studies		14.1

4) Personal Development		*date*	*cp*
►	**General skill training courses**		
	WGS PhD Workshop Carousel	24 May 2019	0.3
	WGS PhD Workshop Carousel - Online	21-27 May 2021	0.3
	WGS PhD Workshop Carousel	12 May 2023	0.2
	WGS PhD Competence Assessment	25 Sep 2019	0.3
	WGS Course: Scientific Artwork, Data visualisation and Infographics with Adobe Illustrator	08-09 Jun 2020	0.6
	WGS Course: Brain Training	16 Jun 2020	0.3
	EPS Introduction Course	28 Sep 2020	0.3
	Wageningen in'to Languages Course: Improve your English: a self-help guide	5 Oct 2020	0.1
	WGS workshop: Critical Thinking and Argumentation	15 Oct 2020	0.3
	WGS Course: Research Data Management part 1-3	12-26 Nov 2020	0.5
	WGS Course: Supervising BSc and MSc thesis students	14-15 Dec 2020	0.6
	WGS Workshop: Reviewing a Scientific Manuscript	5 Feb 2021	0.1
	WGS Course: Mobilising your Scientific Network	23 & 30 Sep 2021	1.0
	The Good Lobby: Lobbying Workshop, hosted by GeneSprout Inititative	23 Apr 2022	0.2
	EPS workshop: Applying for Marie Sklodowska-Curie Fellowship: from proposal to project	22 Jun 2022	0.2
	WGS Workshop: Last Stretch of the PhD program	25 Nov 2022	0.0
►	**Organisation of meetings, PhD courses or outreach activities**		
	GeneSprout Initiative Activities		
	Organisation & Presentations at the following events		
	Gene editing in plantenveredeling @ Jonge Democraten (Wageningen)	29 Oct 2019	0.1
	Social Entrepreneurship: The GeneSprout Initiative @ Current Topics in Biotechnology (WUR)	24 Nov 2019	0.1
	The GeneSprout Initiative @ Slow Food Youth Network event (WUR Fieldlab, Lelystad)	05 Sep 2020	0.1
	Genome Editing in Europe: a Young Researchers Perspective @ EuropaBio (online)	29 Sep 2020	0.1
	Gene Editing in Plant Breeding @ Current Topics in Biotechnology (WUR, online)	01 Oct 2020	0.1
	The Future of Tomorrow: CRISPR and Gene Editing Roundtable Discussion @ CHIC Project (WUR, online)	27 Nov 2020	0.1
	Plant Breeding and its History: Use cases of Precision Breeding @ Plantae (online)	07 Dec 2020	0.1
	GMO Discussion @ Wageningen Student Farm	12 May 2021	0.1
	EPS CRISPR/Cas Workshop: the GeneSprout Inititative	14 Sep 2021	0.1
	Master Class in Seed Technology: CRISPR, NGTs and the GeneSprout Initiative	21 Oct 2021	0.1
	Organising the following events		
	CRISPRcon 2019 side event: GeneSprout Initiative (Wageningen)	Apr-Jun 2019	0.5
	EuropaBio biotech week (Brussels)	23 Sep 2019	0.2
	EuropaBio roundtable event (online)	29 Sep 2020	0.3
	Moderating panel discussion at the following event		
	Plants & Patents 2022 (Wageningen, hybrid)	15 Jun 2022	1.0
►	**Membership of EPS PhD Council**		
	Subtotal Personal Development		8.3

TOTAL NUMBER OF CREDIT POINTS*	**48.0**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was performed at the Laboratory of Biochemistry at Wageningen University & Research and financially supported by a grant from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO; GSGT.GSGT.2018.013) and by Plantum, Rijk Zwaan B.V. and Royal van Zanten B.V.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Fenna Schaap (www.fennaschaap.nl), cover images by Juriaan Rienstra

Thesis layout by Nils Rotgans and Juriaan Rienstra

Printed by ProefschriftMaken.nl

