

Evolutionary analysis of a billion years of auxin biology

Sumanth Kumar Mutte

Propositions

1. All components of ‘modern’ auxin biology originated in the ancestors of land plants.

(this thesis)

2. Despite the lack of key protein domains, non-canonical components are important for diversity and specificity in auxin biology.

(this thesis)

3. ‘Not all disabilities are immediately visible’ - applies not only to priority seats in public transport but also to mutant phenotypes in the plant sciences.

4. Generating data is not the way forward, extensive interpretation with minimal testing is the key.

5. Low expectations and little deliberation are the key factors of success in ‘arranged’ marriages.

6. Respect and trust of humans are taken over by fear and control.

Propositions belonging to the thesis, entitled
Evolutionary analysis of a billion years of auxin biology

Sumanth Kumar Mutte
Wageningen, 15 April 2020

Evolutionary analysis of a billion years of auxin biology

Sumanth Kumar Mutte

Thesis committee

Promotor

Prof. Dr D. Weijers
Professor of Biochemistry
Wageningen University and Research

Other members

Prof. Dr B.J. Zwaan, Wageningen University and Research
Prof. Dr B. Snel, Utrecht University
Dr F. Parcy, CEA Grenoble, France
Dr F.V. Mironova, Novosibirsk State University, Russian Federation

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

Evolutionary analysis of a billion years of auxin biology

Sumanth Kumar Mutte

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University

by the authority of the Rector Magnificus

Prof. Dr A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Wednesday 15 April 2020

at 4 p.m. in the Aula.

Sumanth Kumar Mutte

Evolutionary analysis of a billion years of auxin biology

142 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2020)

With references, with summary in English

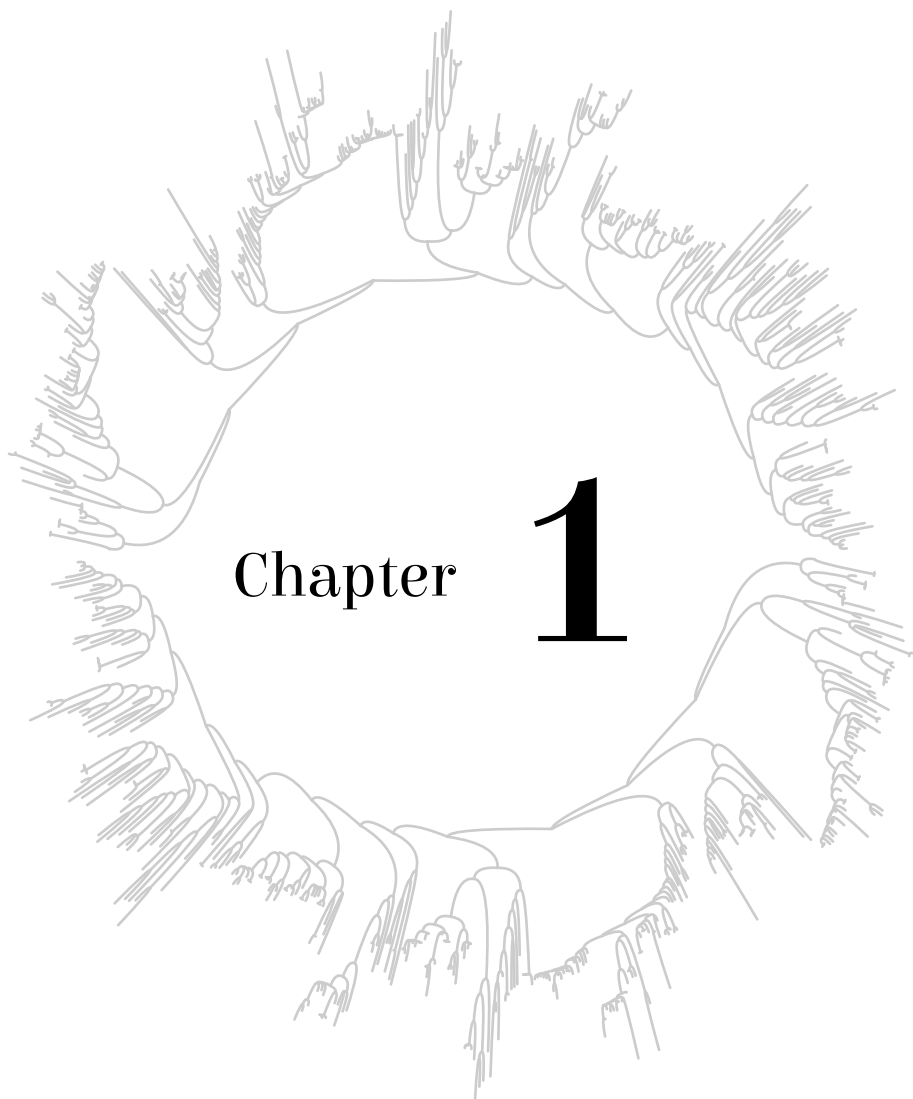
ISBN: 978-94-6395-269-9

DOI: 10.18174/511001

నా తల్లిదండ్రులు మంజుల మరియు విడుకొండలు కు అంకితం ...

Table of Contents

Chapter 1	8
Introduction	
Chapter 2	24
High-resolution and deep phylogenetic reconstruction of ancestral states from large transcriptomic data sets	
Chapter 3	40
Reconstructing the evolutionary past of auxin biology	
Chapter 4	60
Origin and evolution of the nuclear auxin response system	
Chapter 5	96
Deep evolutionary history of the Phox and Bem1 (PB1) domain across eukaryotes	
Chapter 6	124
General discussion	
English Summary	134
Acknowledgements	136
Curriculum Vitae	138
Publications	139
Education Statement	140



Chapter 1

Introduction

Origin of life

The emergence of life is undoubtedly one of the most important events in the history of earth since its origin around 4 billion years ago. Since then, there has been a substantial diversity in the evolution of life ranging from simple unicellular forms to the complex multicellular organisms that are either extinct or extant. Unicellular prokaryotes are the most primitive organisms on earth, and formed the basis for evolution of life and subsequently the more advanced form of life, the eukaryotes. In the last few decades, Archaea were found to be the closely related to eukaryotes (Woese and Fox, 1977; Fig. 1). Based on the current molecular phylogenetics data, the superphylum Asgard archaea is identified as the sister group to eukaryotes, being more closely related to eukaryotes than to bacteria (Woese et al., 1990). Hence, the Last Eukaryote Common Ancestor (LECA), that evolved around ~1-1.9 billion years ago, is presumed to contain the features of both bacterial and archaeal origin (Eme et al., 2014).

LECA is the most primitive eukaryote, and has been inferred to have a nucleus containing linear chromosomes, as well as elaborate gene expression and protein regulation systems. It is concluded to have obtained mitochondria from the endosymbiosis of free-living alphaproteobacteria (Gray, 2012; Koumandou et al., 2013). LECA gave rise to a diverse set of both unicellular and multicellular eukaryotes, that are divided into five kingdoms: Protozoa, Chromista, Plantae (Archaeplastida), Fungi and Animalia (Metazoa; Fig.1; Adl et al., 2012; Ruggiero et al., 2015). Even though the kingdom-based classification of eukaryotes is widely accepted, studies based on molecular phylogenetics have identified more genera in the recent years. This has led to new division into six or seven supergroups (Adl et al., 2019; Burki et al., 2019; Keeling et al., 2005; Simpson and Roger, 2004). The kingdom Chromista is now divided into three supergroups: TSAR, Haptista and Cryptista. Amoebozoa along with Fungi and Animalia are placed in the supergroup Amorphea, whereas other Protozoans are spread across the supergroups CRuMs and Excavata. The kingdom Plantae or Archaeplastida is considered as a supergroup in itself (Burki et al., 2019).

Evolution of plants

Archaeplastida presumably originated more than a billion years ago with the acquisition of a plastid or chloroplast through the endosymbiosis of Cyanobacteria (Delwiche et al., 1995). The monophyletic Archaeplastida is comprised of three main groups of organisms: Rhodophyta (red algae), Glaucophyta and Viridiplantae (green algae and land plants; Fig. 2; Kenrick and Crane, 1997). In nature, glaucophytes are not as diverse and species-rich as red algae or the land plants (Delaux et al., 2012a). Viridiplantae are significantly important in the biosphere as they contain chlorophyll, a photosynthetic pigment that is needed for harnessing the energy in light, which ultimately leads to starch synthesis in chloroplasts. Green algae are further divided into chlorophytes and charophytes. Chlorophytes are mostly marine unicellular organisms, with independent and multiple origins of multicellularity identified in various clades (Delaux et al., 2012a; Marin, 2012). Conversely, the majority of the charophytes are freshwater organisms

and some species live even in the subaerial or subterrestrial environments (Cheng et al., 2019). Charophytes are divided into Mesostigmatophyceae, Klebsormidiophyceae, Charophyceae and Zygnematophyceae, among other classes. Recently sequenced genomes of these classes revealed that many gene families that were previously thought to be limited to land plants were in fact also found in charophytes (Cheng et al., 2019; Hori et al., 2014; Liang et al., 2019; Nishiyama et al., 2018). Interestingly, recent molecular phylogenetic studies placed Zygnematophyceae as the sister clade to the land plants. These algae are adapted to subaerial or subterrestrial environment despite their simple morphology, when compared to Charophyceae (Cheng et al., 2019; Nishiyama et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019; Timme and Delwiche, 2010).

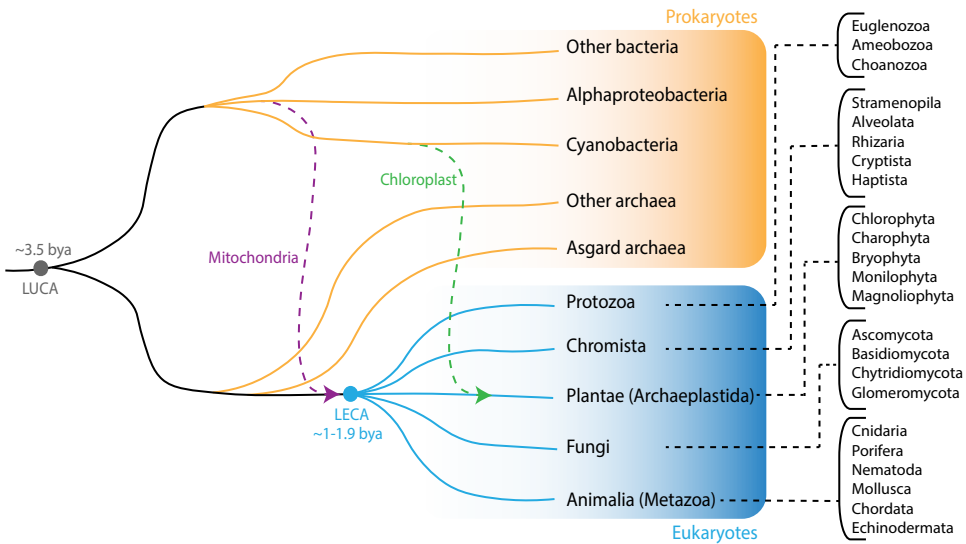


Figure 1: Simple schematic showing major forms of life on earth. Prokaryotes (bacteria and archaea) are indicated in orange and eukaryotes are indicated in blue. Asgard archaea is the closest superphylum identified so far as a sister clade to the Last Eukaryote Common Ancestor (LECA). LECA has acquired mitochondria by the endosymbiosis of an alphaproteobacterium. Chloroplast in the kingdom Plantae is obtained by the endosymbiosis of a cyanobacterium in the common ancestor of Archaeplastida. Various well studied (super)phyla from each eukaryotic kingdom are indicated on the right with corresponding dotted lines. bya, billion years ago.

Along with the transition from aquatic-to-terrestrial habitats, plants underwent both morphological and physiological transformations, with specialized mechanisms to control water loss and gas exchange (i.e., cuticle and stomata), structures for extracting moisture below the soil surface (i.e., roots) and tolerance to desiccation (Delaux et al., 2012a; Graham, 1993). Another key transformation is the shift from haploid-dominant life cycle in green algae to diploid-dominant life cycle in land plants (Haig, 2010). Bryophytes are the early diverged land plants, with hornworts as its basal lineage, being sister clade to mosses and liverworts (Fig. 2; Puttick et al., 2018; Wickett et al., 2014). Bryophytes do not contain either vascular tissues (xylem and phloem) or true roots. Instead they contain root hair-like structures called rhizoids. Early diverged vascular plants are the lycophytes, which is a sister group for euphyllophytes (ferns

and seed plants; Fig. 2). Vascular plants are characterized by the development of well-defined conducting vascular tissues. With the origin of vascular plants, the species have witnessed the expansion of various gene families, which further expanded in the later diverged seed plants or gymnosperms (One Thousand Plant Transcriptomes Initiative, 2019). Later in evolution, flowering plants (angiosperms) has become by far the most species-rich and highly diverged land plant group with nearly 370,000 species (Lughadha et al., 2016). Despite their highly complex biochemical, morphological and physiological innovations, there seems to be a restriction to the number of gene families, which can be attributed to a process of gene co-option, where the new processes are controlled by already existing genes in the seed plant ancestors (Amborella Genome Project, 2013; One Thousand Plant Transcriptomes Initiative, 2019).

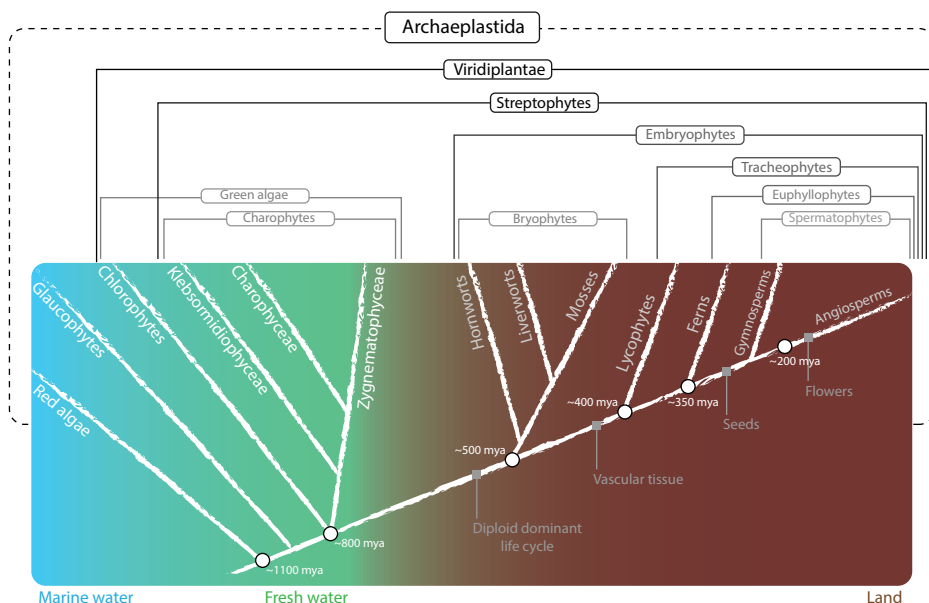


Figure 2: Evolution of Archaeplastida. Glaucophytes are basal to all green algae and land plants (Viridiplantae). Bryophytes are the early diverged land plants, that further evolved into vascular plants (Tracheophytes) and seed plants (Spermatophytes). Various commonly used grouping conventions are indicated in the top with half-rectangles. Black circles indicate the key evolutionary points with time rounded-off time estimates from Morris et al., 2018. mya, million years ago.

Phytohormones

Plant hormones such as auxin, cytokinin (CK), ethylene, jasmonic acid (JA), gibberellic acid (GA), abscisic acid (ABA) and strigolactones (SL) are crucial for plant growth at all developmental stages. Many aspects of phytohormone synthesis, signalling, transport and metabolism have been relatively well established (reviewed in Santner and Estelle, 2009). JA protects plants against wounding, herbivores and some pathogens, whereas CKs influence cell division, chloroplast development, leaf senescence, and root or shoot branching (Davies, 2010). Auxin plays a central role in plant growth and development by affecting cell expansion, division and differentiation.

Several plant hormone pathways that were previously thought to be specific to land plants, were later found even in charophyte algae (Hori et al., 2014; Nishiyama et al., 2018). Auxin-dependent responses and transport were observed in *Klebsormidium nitens* (Ohtaka et al., 2017; Skokan et al., 2019), indicating the existence of auxin transport in ancestors of charophytes. SL modulate seed germination and photomorphogenesis and also inhibit root and shoot branching in angiosperms, whereas it stimulates the elongation of rhizoids in charophytes, liverworts and mosses (Delaux et al., 2012b; Proust et al., 2011).

Ethylene regulates both development and defence processes of plants in response to (a)biotic stress. Ethylene biosynthesis and responses were observed to be deeply conserved even in charophytes, analogous to angiosperms (Ju et al., 2015). ABA plays an important role in response to environmental stresses, especially drought. Recent studies showed that, despite the presence of homologs in charophytes for ABA regulation, hormonal control is attained by the co-option of pre-existing response components in vascular plants (Sun et al., 2019). In a similar way, despite the presence of some components of the GA pathway in basal land plants, a complete GA pathway is limited to the vascular plants (Hernández-García et al., 2019; Hirano et al., 2007). Hence, there is a clear step-wise evolution of various hormone signaling components in different pathways at multiple stages of Viridiplantae evolution. However, the majority of these studies are limited to one or few genes within a pathway, and inferences are generally based on only few species. Thus, models for plant hormone response evolution may not represent the entire hormone synthesis and regulation across all phyla in land plants. Hence, we need a deeper understanding of pathway evolution (both synthesis and signaling), along with the detailed divergence of each ortholog that we see today in extant land plant phyla.

Auxin: Synthesis to signalling

Auxin is a key phytohormone that regulates various aspects of plant growth and development. Indole-3-acetic acid (IAA) is a naturally synthesized auxin in plants. Among the various possible routes of IAA synthesis proposed so far (Brumos et al., 2014; Zhao, 2014), the two-step conversion of Tryptophan (Trp) to IAA through Indole-3-pyruvic acid (IPA) is the most studied and most well-established route of biosynthesis (Fig. 3; Zhao, 2012). The key enzymes involved in this conversion are TRYPTOPHAN AMINOTRANSFERASE OF ARABIDOPSIS (TAA) aminotransferases and the YUCCA (YUC) family of flavin-containing monooxygenases. TAA converts Trp to IPA, which is then converted to IAA by YUC proteins (Fig. 3; Zhao, 2012). While alternative routes for auxin biosynthesis have been reported – involving CYP79B2/CYP79B3 (Zhao, 2002), aldehyde oxidase (Seo et al., 1998), and IPA decarboxylase (Vande Broek et al., 1999) – it is important to note that the TAA/YUC pathway appears to be the major one. Genetic analysis of mutants in TAA and the related TAA-RELATED 1 and TAA-RELATED 2 (TAR1 and TAR2) genes cause strong growth and developmental defects (Stepanova et al., 2008). Likewise, higher-order mutants that knock out several members of the YUC family cause similar, severe developmental defects (Cheng et al., 2006, 2007). No such strong phenotypes were reported in

mutants in alternative auxin synthesis pathways (Mashiguchi et al., 2011; Normanly et al., 1997; Zhao, 2002), which suggests that the TAA/YUC pathway provides a critical, if not the major source of auxin during normal development.

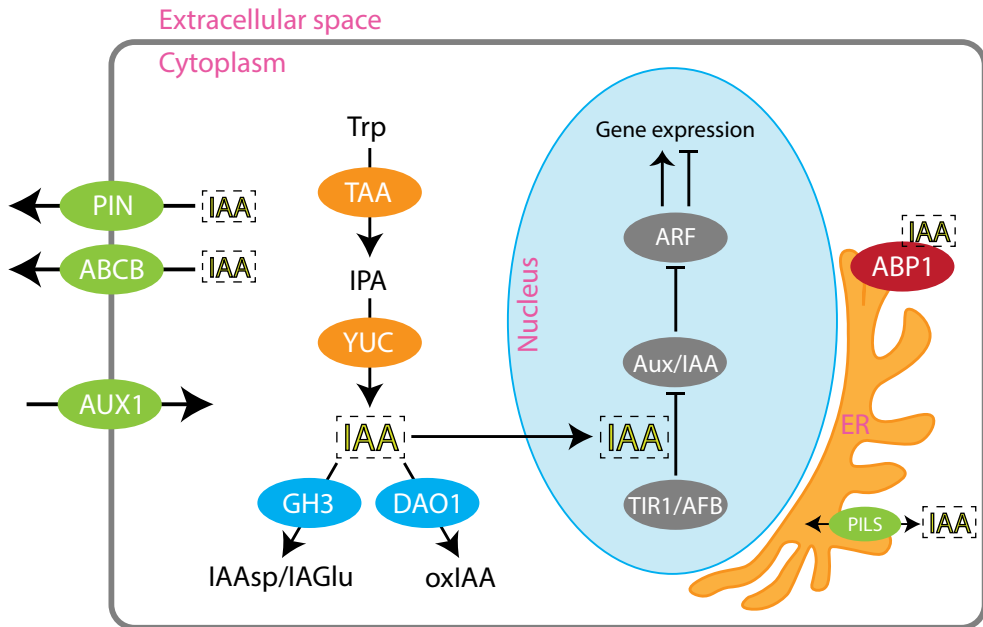


Figure 3: Simplified overview of the auxin pathway in plants. Proteins indicated in green ovals (PIN, ABCB and AUX1) represent the auxin transporters, orange ovals (TAA and YUC) are the biosynthesis genes, blue ovals (GH3 and DAO1) are the auxin metabolic families, NAP components localized in the nucleus (ARF, Aux/IAA and TIR1/AFB) are indicated in grey and non-genomic pathway component (ABP1) in red, localized in the endoplasmic reticulum (ER). PIN: PIN-FORMED; PILS: PIN-LIKES; ABCB: ATP-BINDING CASSETTE SUBFAMILY B; AUX1: AUXIN RESISTANT 1; TAA: TRYPTOPHAN AMINOTRANSFERASE OF ARABIDOPSIS; YUC: YUCCA; GH3: GRETCHEN HAGEN 3; DAO1: DIOXYGENASE FOR AUXIN OXIDATION 1; ARF: AUXIN RESPONSE FACTOR; Aux/IAA: AUX-IN/INDOLE-3-ACETIC ACID; TIR1/AFB: TRANSPORT INHIBITOR RESPONSE 1/AUXIN SIGNALING F-BOX.


It has been recognized long ago that auxin can be transported from its site of application or synthesis (Goldsmith, 1977; Rubery and Sheldrake, 1974). Decades of physiological, biochemical and genetic research has identified a well-supported model for directional auxin transport. Various membrane proteins have been shown or suggested to transport auxin, or to facilitate its transport across membranes (Bennett et al., 1996; Chen et al., 2001; Galweiler et al., 1998; Jahrmann et al., 2005). Among these proteins, PIN-FORMED (PIN) proteins have been studied in most detail. The first PIN protein was identified through genetic analysis of the *pin1* mutant (Okada et al., 1991), a classical auxin transport mutant that lacks flowers, in *Arabidopsis*. The PIN1 gene was found to encode a protein with multiple transmembrane helices, superficially similar to bacterial membrane-resident transporter proteins. The most striking finding was that PIN1 protein was polarly localized in the membrane, facing in the direction of polar auxin transport (Galweiler et al., 1998). Later, other members of the PIN

family in *Arabidopsis* were shown to polarly localize and act in embryogenesis, (lateral) root formation, tropisms and patterning (Benková et al., 2003; Blilou et al., 2005; Friml et al., 2002; Weijers et al., 2005). While direct transporter activity has not been formally shown, it is clear that PIN proteins facilitate selective auxin transport. PIN proteins are functionally conserved in various organisms, including charophytes (Benková et al., 2003; Friml et al., 2002; Skokan et al., 2019; Viaene et al., 2013). Later, another auxin transport facilitator family homologous to PIN proteins, PIN-LIKES (PILS), was identified to be localized in the endoplasmic reticulum (ER) and regulates intracellular auxin homeostasis in plants (Barbez et al., 2012). Another family of efflux transporters are the ATP-BINDING CASSETTE SUBFAMILY B/MULTIDRUG RESISTANCE/PHOSPHOGLYCOPROTEIN (ABCB/MDR/PGP) protein family (Noh et al., 2001; Verrier et al., 2008). Similar to PIN proteins, ABCB family orthologs are also involved in root elongation, lateral root formation, gravitropism and embryogenesis (Mravec et al., 2008). However, these transporters appear to serve an accessory role, and their localization is not polarized. Conversely, another family of auxin transporters were identified that belong to the AUXIN RESISTANT 1/LIKE AUX1 (AUX1/LAX) family of transmembrane proteins. These are proton gradient-driven influx transporters, and are involved in promoting lateral root emergence (Bennett et al., 1996; Swarup and Péret, 2012) as well as in various other developmental processes, including embryogenesis (Robert et al., 2015; Ugartechea-Chirino et al., 2010).

While the transport of auxin is mediated by the influx and efflux carriers, the direction and kinetics of transport is actively regulated during plant development. An important aspect of transport regulation is the abundance of transporters on the plasma membrane, a process that is actively controlled by regulated endocytosis and exocytosis of vesicles containing the transporter proteins (Žárský et al., 2009). Various proteins have been identified that control either the activity of the PIN proteins (e.g. protein kinases; Christensen et al., 2000), or their abundance at the plasma membrane (Luschnig and Vert, 2014).

Apart from transport, another mechanism to maintain cellular auxin levels or homeostasis is oxidation or conjugation (Ljung, 2013). IAA is irreversibly converted to Oxindole-3-acetic acid (oxIAA) by the DIOXYGENASE FOR AUXIN OXIDATION 1 (DAO1) enzyme (Fig. 3). Mutants in this enzyme showed increased lateral root density, among other morphophysiological changes (Zhang et al., 2016). Furthermore, conjugation of IAA to amino acids represents a reversible inactivation process, hence considered as a storage mechanism (Ludwig-Müller, 2011; Zhang et al., 2017). IAA is converted to amide conjugates such as Indole-3-acetyl-aspartic acid (IAA_{asp}) and Indole-3-acetyl-glutamic acid (IAA_{glu}) by a family of GRETCHEN HAGEN 3 (GH3) amide synthases (Staswick et al. 2005; Fig. 3). Interestingly, some GH3 paralogs also have preference not only for auxin, but also for other hormones such as jasmonate and salicylic acid (Nobuta et al., 2007; Okrent et al., 2009; Staswick et al., 2002).

Plants respond in various ways to treatment with auxin. Notably, auxin treatment inhibits root and shoot growth in *Arabidopsis* seedlings, and this phenotype has been an



important starting point for genetic analysis of auxin response (Abel et al., 1995). Through isolating auxin-resistant mutants, the core of a nuclear response system was identified: the Nuclear Auxin Pathway (NAP). The key components in this pathway are the auxin receptors TRANSPORT INHIBITOR RESPONSE 1/AUXIN SIGNALLING F-BOX (TIR1/AFB) proteins (Dharmasiri et al., 2005a). IAA binds to these F-box proteins and acts as molecular glue in forming a co-receptor complex with AUXIN/INDOLE-3-ACETIC ACID (Aux/IAA) transcriptional repressor proteins (Dharmasiri et al., 2005b). At low auxin concentrations, Aux/IAA proteins bind to and inhibit AUXIN RESPONSE FACTORS (ARF), the transcription factors that regulate auxin response genes (Tiwari et al., 2003, 2004). At high IAA concentrations, the Aux/IAA proteins are subjected to ubiquitin-dependent proteasomal degradation, releasing ARF proteins for downstream gene regulation (Calderon-Villalobos et al., 2010). Hence, ARF proteins in NAP regulate thousands of downstream targets and control a plethora of processes in plant growth and development (reviewed in Weijers and Wagner, 2016). Importantly, genetic analysis suggests that this pathway is key to auxin-dependent development: higher-order mutations in multiple TIR1/AFB family members cause phenotypes that are very similar to those found in auxin-deficient *taa1* or *yuc* mutants (Dharmasiri et al., 2005a; Parry et al., 2009).

An alternate, proteasome-independent and non-genomic auxin response pathway has emerged where another protein, AUXIN BINDING PROTEIN 1 (ABP1; Woo, 2002) was identified to bind auxin. This protein was suggested to play an important role in cell division and cell elongation (Sauer and Kleine-Vehn, 2011), presumably in part by controlling endocytosis of PIN proteins (Robert et al., 2010). However, more recent analysis of CRISPR/Cas-induced loss of function mutants in *Arabidopsis* ABP1 suggest that ABP1 is not required for auxin signaling or *Arabidopsis* development, making its role in auxin responses and plant growth still uncertain (Gao et al., 2015; Paponov et al., 2019). Recent studies have indicated that there could be even quick non-transcriptional auxin response mechanism through TIR1 dependent signaling (Fendrych et al., 2018) or an alternate TRANSMEMBRANE KINASE 1 (TMK1) dependent non-canonical auxin response (Cao et al., 2019). However, it is still unclear if there are any other auxin receptors that sense auxin or translate auxin signals quicker than TIR1/AFB or ABP1.

Auxin dependent responses – Underlying similarities

Auxin is important for morphogenesis and establishment of tissues in angiosperms, which have high genetic redundancy when compared to more early diverged species, such as those in the bryophytes. Recent studies in the model moss *Physcomitrella patens* revealed that components and functionality of the NAP are conserved with role in *Arabidopsis*. By studying mutants and knock-outs of Aux/IAA genes, it was shown that Aux/IAA's in *Physcomitrella* act in a degradation- and ARF-dependent pathway to control development (Lavy et al., 2016; Prigge et al., 2010). In *Marchantia polymorpha*, a model liverwort, transgenic plants expressing a non-degradable Aux/IAA mutants showed auxin-insensitive phenotype (Kato et al., 2015). These studies indicate that bryophytes and angiosperms share a common auxin perception system. ARFs are classified


into three classes A, B and C, where the class-A and class-B function as transcriptional activators and repressors, respectively (Ulmasov et al., 1999). Interestingly the respective orthologs of class-A and class-B ARFs in both *M. polymorpha* and *P. patens* also showed similar transactivation patterns (Kato et al., 2015; Lavy et al., 2016). Moreover, canonical auxin biosynthesis and PIN-dependent auxin transport are also needed for development of these two species (Bennett et al., 2014; Eklund et al., 2015; Viaene et al., 2013). This indicates that three major aspects of auxin biology - biosynthesis, transport and signalling - are all conserved between bryophytes and angiosperms. Recently published charophyte genomes of *Klebsormidium nitens* and *Chara braunii*, indicated that not all the components of auxin pathway are encoded in these genomes (Hori et al., 2014; Nishiyama et al., 2018).

In *Arabidopsis*, on the exogenous application of auxin, many mutants of the NAP genes show defects in cell expansion and tropic responses, where similar effects were observed in *Marchantia* (Flores-Sandoval et al., 2015, 2018; Kato et al., 2015). It was also shown that the NAP is critical for axis formation during embryogenesis in *Arabidopsis* (Hardtke and Berleth, 1998; Weijers et al., 2006). Likewise, gemmae development in *Marchantia* depends on normal ARF activity: a mutation in the class-A ARF1 in *Marchantia* leads to characteristic division defects in young vegetative propagules (gemmae; Kato et al., 2015). Mutations in the *Arabidopsis* ARF5/MONOPTEROS (MP; class-A ARF) cause at least superficially similar defects in cell division in the young embryo (Hardtke and Berleth, 1998). An interesting question therefore is whether common mechanisms control auxin-dependent cell division patterns in these two systems (reviewed in Kato et al., 2018). In *Arabidopsis*, class II RSL genes that are important for root hair development, are conserved in *M. polymorpha* and shown to be involved in rhizoid formation (Menand et al., 2007; Proust et al., 2016). However, whether they are regulated by auxin in *M. polymorpha* is still unclear. In *M. polymorpha*, class-A ARF functions as a positive regulator of gemmae dormancy, similar to seed dormancy in spermatophytes (Eklund et al., 2015).

Since we now know that both early and recently diverged land plants, bryophytes and angiosperms, have conserved auxin-dependent regulatory and developmental patterns, it is important to understand how the step-wise complexity in each gene family has led to the step-wise morpho-physiological innovations in the land plants i.e. vasculature in tracheophytes, seeds in gymnosperms and flowers in angiosperms. It is important to underpin the causality and correlations of the novel gene copies (orthologs) to novel features in land plant evolution.

Scope of this thesis

Phylogenies enrich our understanding of genes, genomes, pathways, systems and organisms. They help us understand the patterns and relations between different systems and help us find the correlations between mechanisms to the relations between organisms. With the growth of 'omics' data and advancement of new tools and technologies in molecular phylogenetics, these correlations are becoming ever more accurate. Whereas earlier studies were limited by the



genome sequence data of few model organisms, the advancement of next generation (or high-throughput) sequencing and transcriptomics has rapidly increased the availability of gene content of various non-model organisms across all the kingdoms of life. Even though transcriptome data is inherently limited and not as complete and detailed as genome information, studying homologs and evolutionary patterns between genes and underlying gene family expansions is possible. However, there is a possibility that the gene or gene family under investigation is not expressed in the tissue or the species. In this case, data from multiple tissues and organisms that belong to the same family or class should help in understanding the ancestral state i.e. the minimal gene complement of that gene in that particular lineage. However, due to the inherent limitations of the transcriptome data, species-specific gene duplications and gene losses are not estimated which can only be obtained from genome information.

In the last decade, plant sciences have seen an enormous increase in molecular phylogenetics studies, largely due to the availability of more than 1300 transcriptomes from all the major classes in the kingdom Plantae. By taking advantage of this enormous dataset we have developed a simple yet effective methodology in **Chapter 2** to reconstruct the origin and evolution of various genes families across Archaeplastida.

We first used the methodology described in Chapter 2 to study the evolution of auxin biosynthesis gene families TAA and YUC and a major gene family responsible for auxin homeostasis i.e. the GH3 in **Chapter 3**. As auxin can elicit both genomic and non-genomic responses, evolution of one of the widely known non-genomic auxin pathway component, ABP1, was also studied in detail. Finally, we briefly looked at the evolution of components of downstream auxin responses in at least two processes i.e. vascular development (TMO5/LHW) and cell polarity (SOK). The SOK gene family is not functionally annotated, hence we also performed the motif/domain annotation along with deep evolution across eukaryotes.

In **Chapter 4**, we focus on the major auxin signalling and genomic response pathway. We study both the origin and evolution of NAP components, ARF, Aux/IAA and TIR1/AFB in detail. Further, we study the early auxin response capacity of some species in early diverged lineages (hornworts, liverworts and mosses) and the vascular plants (ferns), by comparing the auxin treated transcriptome to the untreated plants using RNA-Seq. We also included two charophyte species in this study. As there are genomes available only for model liverwort (*Marchantia*) and moss (*Physcomitrella*), but not others, we built de novo transcriptomes and performed differential expression analysis. Further, we studied the core auxin response gene set that might indicate the common responses of not only the pathways but also the downstream components. As a surprise, we identified deeply conserved non-canonical components, which we tested for their role in auxin dependent responses by studying the mutants in *Marchantia*.

A key module in the function of NAP is the interaction between ARF and Aux/IAA proteins, mediated by the C-terminal Phox and Bem1 (PB1) domain. As PB1 domain is also identified in Animals as well as in Fungi, we studied in **Chapter 5** if the PB1 domain originated in LECA. Moreover, we also investigated if there are other gene families in plants that possess a

PB1 domain. We have also generated amino acid descriptor based Random Forest classification to differentiate various PB1 domains across land plants. We further studied the correlation of these descriptors using homology modeling.

Chapter 6 summarizes and discusses the most important findings of this thesis and provides direction for future research.

Acknowledgement

We thank Shubhajit Das and Dr. Simon Lindhoud for helpful comments on the manuscript.

References

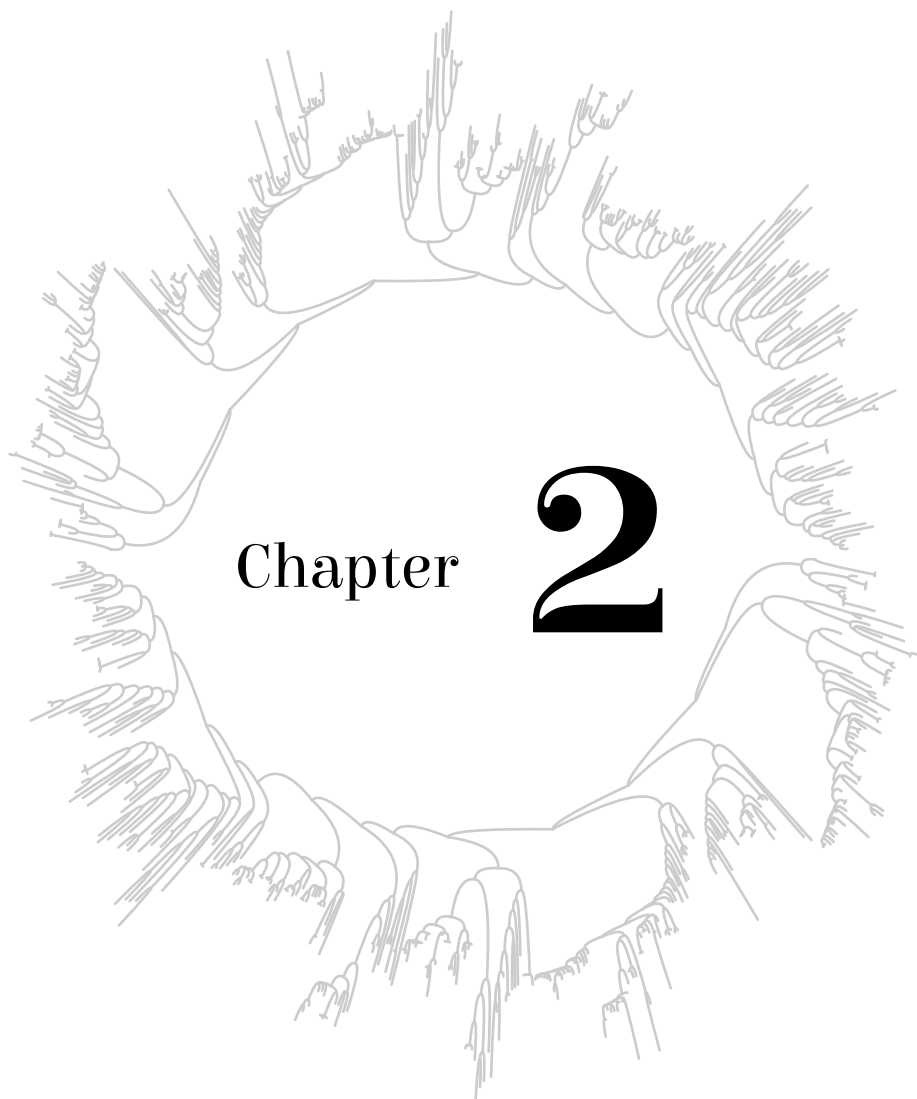
- Abel, S., Nguyen, M.D., and Theologis, A. (1995). ThePS-IAA4/5-like Family of Early Auxin-inducible mRNAs in *Arabidopsis thaliana*. *J. Mol. Biol.* 251, 533–549.
- Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., et al. (2012). The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* 59, 429–493.
- Adl, S.M., Bass, D., Lane, C.E., Lukeš, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F., et al. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *J. Eukaryot. Microbiol.* 66, 4–119.
- Amborella Genome Project (2013). The Amborella Genome and the Evolution of Flowering Plants. *Science* 342, 1241089.
- Barbez, E., Kubeš, M., Rolčík, J., Béziat, C., Pěňčík, A., Wang, B., Rosquete, M.R., Zhu, J., Dobrev, P.I., Lee, Y., et al. (2012). A novel putative auxin carrier family regulates intracellular auxin homeostasis in plants. *Nature* 485, 119–122.
- Benková, E., Michniewicz, M., Sauer, M., Teichmann, T., Seifertová, D., Jürgens, G., and Friml, J. (2003). Local, Efflux-Dependent Auxin Gradients as a Common Module for Plant Organ Formation. *Cell* 115, 591–602.
- Bennett, M.J., Marchant, A., Green, H.G., May, S.T., Ward, S.P., Millner, P.A., Walker, A.R., Schulz, B., and Feldmann, K.A. (1996). *Arabidopsis* AUX1 Gene: A Permease-Like Regulator of Root Gravitropism. *Science* 273, 948 – 950.
- Bennett, T., Brockington, S.F., Rothfels, C., Graham, S.W., Stevenson, D., Kutchan, T., Rolf, M., Thomas, P., Wong, G.K.S., Leyser, O., et al. (2014). Paralogous radiations of PIN proteins with multiple origins of noncanonical PIN structure. *Mol. Biol. Evol.* 31, 2042–2060.
- Blilou, I., Xu, J., Wildwater, M., Willemssen, V., Paponov, I., Friml, J., Heidstra, R., Aida, M., Palme, K., and Scheres, B. (2005). The PIN auxin efflux facilitator network controls growth and patterning in *Arabidopsis* roots. *Nature* 433, 39–44.
- van de Broek, A., Lambrecht, M., Eggermont, K., and Vanderleyden, J. (1999). Auxins upregulate expression of the indole-3-pyruvate decarboxylase gene in *Azospirillum brasilense*. *J. Bacteriol.* 181, 1338–1342.
- Brumos, J., Alonso, J.M., and Stepanova, A.N. (2014). Genetic aspects of auxin biosynthesis and its regulation. *Physiol. Plant.* 151, 3–12.
- Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2019). The New Tree of Eukaryotes. *Trends Ecol. Evol.* In press, 1–13.
- Calderon-Villalobos, L.I., Tan, X., Zheng, N., and Estelle, M. (2010). Auxin perception--structural insights. *Cold Spring Harb. Perspect. Biol.* 2, a005546.
- Cao, M., Chen, R., Li, P., Yu, Y., Zheng, R., Ge, D., Zheng, W., Wang, X., Gu, Y., Gelová, Z., et al. (2019). TMK1-mediated auxin signalling regulates differential growth of the apical hook. *Nature* 568, 240–243.
- Chen, L., Ortiz-Lopez, A., Jung, A., and Bush, D.R. (2001). ANT1, an Aromatic and Neutral Amino Acid Transporter in *Arabidopsis*. *Plant Physiol.* 125, 1813–1820.
- Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T., Sun, W., Li, X., Xu, Y., Zhang, Y., et al. (2019). Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* 179, 1057-1067.e14.
- Cheng, Y., Dai, X., and Zhao, Y. (2006). Auxin biosynthesis by the YUCCA flavin monooxygenases controls the formation of floral organs and vascular tissues in *Arabidopsis*. *Genes Dev.* 20, 1790–1799.
- Cheng, Y., Dai, X., and Zhao, Y. (2007). Auxin Synthesized by the YUCCA Flavins Monooxygenases Is Essential for Embryogenesis and Leaf Formation in *Arabidopsis*. *Plant Cell* 19, 2430–2439.

- Christensen, S.K., Dagenais, N., Chory, J., and Weigel, D. (2000). Regulation of Auxin Response by the Protein Kinase PINOID. *Cell* 100, 469–478.
- Davies, P.J. (2010). The Plant Hormones: Their Nature, Occurrence, and Functions. In *Plant Hormones: Biosynthesis, Signal Transduction, Action!*, P.J. Davies, ed. (Dordrecht: Springer Netherlands), pp. 1–15.
- Delaux, P.-M., Nanda, A.K., Mathé, C., Séjalon-Delmas, N., and Dunand, C. (2012a). Molecular and biochemical aspects of plant terrestrialization. *Perspect. Plant Ecol. Evol. Syst.* 14, 49–59.
- Delaux, P.M., Xie, X., Timme, R.E., Puech-Pages, V., Dunand, C., Lecompte, E., Delwiche, C.F., Yoneyama, K., Bécard, G., and Séjalon-Delmas, N. (2012b). Origin of strigolactones in the green lineage. *New Phytol.* 195, 857–871.
- Delwiche, C.F., Kuhsel, M., and Palmer, J.D. (1995). Phylogenetic Analysis of *tufA* Sequences Indicates a Cyanobacterial Origin of All Plastids. *Mol. Phylogenet. Evol.* 4, 110–128.
- Dharmasiri, N., Dharmasiri, S., Weijers, D., Lechner, E., Yamada, M., Hobbie, L., Ehrismann, J.S., Jürgens, G., and Estelle, M. (2005a). Plant Development Is Regulated by a Family of Auxin Receptor F Box Proteins. *Dev. Cell* 9, 109–119.
- Dharmasiri, N., Dharmasiri, S., and Estelle, M. (2005b). The F-box protein TIR1 is an auxin receptor. *Nature* 435, 441–445.
- Eklund, D.M., Ishizaki, K., Flores-Sandoval, E., Kikuchi, S., Takebayashi, Y., Tsukamoto, S., Hirakawa, Y., Nonomura, M., Kato, H., Kouno, M., et al. (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.
- Eme, L., Sharpe, S.C., Brown, M.W., and Roger, A.J. (2014). On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. *Cold Spring Harb. Perspect. Biol.* 6, a016139–a016139.
- Fendrych, M., Akhmanova, M., Merrin, J., Glanc, M., Hagihara, S., Takahashi, K., Uchida, N., Torii, K.U., and Friml, J. (2018). Rapid and reversible root growth inhibition by TIR1 auxin signalling. *Nat. Plants* 4, 453–459.
- Flores-Sandoval, E., Eklund, D.M., and Bowman, J.L. (2015). A Simple Auxin Transcriptional Response System Regulates Multiple Morphogenetic Processes in the Liverwort *Marchantia polymorpha*. *PLOS Genet.* 11, e1005207.
- Flores-Sandoval, E., Eklund, D.M., Hong, S.F., Alvarez, J.P., Fisher, T.J., Lampugnani, E.R., Goltz, J.F., Vázquez-Lobo, A., Dierschke, T., Lin, S.S., et al. (2018). Class C ARFs evolved before the origin of land plants and antagonize differentiation and developmental transitions in *Marchantia polymorpha*. *New Phytol.* 218, 1612–1630.
- Friml, J., Wiśniewska, J., Benková, E., Mendgen, K., and Palme, K. (2002). Lateral relocation of auxin efflux regulator PIN3 mediates tropism in *Arabidopsis*. *Nature* 415, 806–809.
- Galweiler, L., Guan, C., Müller, A., Wisman, E., Mendgen, K., Yephremov, A., and Palme, K. (1998). Regulation of Polar Auxin Transport by AtPIN1 in *Arabidopsis* Vascular Tissue. *Science* 282, 2226–2230.
- Gao, Y., Zhang, Y., Zhang, D., Dai, X., Estelle, M., and Zhao, Y. (2015). Auxin binding protein 1 (ABP1) is not required for either auxin signaling or *Arabidopsis* development. *Proc. Natl. Acad. Sci.* 112, 2275–2280.
- Goldsmith, M.H.M. (1977). The Polar Transport of Auxin. *Annu. Rev. Plant Physiol.* 28, 439–478.
- Graham, L.E. (1993). *Origin of land plants*. (New York: John Wiley & Sons, Inc.).
- Gray, M.W. (2012). Mitochondrial Evolution. *Cold Spring Harb. Perspect. Biol.* 4, a011403–a011403.
- Haig, D. (2010). What Do We Know About Charophyte (Streptophyta) Life Cycles? *J. Phycol.* 46, 860–867.
- Hardtke, C.S., and Berleth, T. (1998). The *Arabidopsis* gene *MONOPTEROS* encodes a transcription factor mediating embryo axis formation and vascular development. *EMBO J.* 17, 1405–1411.
- Hernández-García, J., Briones-Moreno, A., Dumas, R., and Blázquez, M.A. (2019). Origin of Gibberellin-Dependent Transcriptional Regulation by Molecular Exploitation of a Transactivation Domain in *della* Proteins. *Mol. Biol. Evol.* 36, 908–918.
- Hirano, K., Nakajima, M., Asano, K., Nishiyama, T., Sakakibara, H., Kojima, M., Katoh, E., Xiang, H., Tanahashi, T., Hasebe, M., et al. (2007). The GID1-Mediated Gibberellin Perception Mechanism Is Conserved in the Lycopphyte *Selaginella moellendorffii* but Not in the Bryophyte *Physcomitrella patens*. *Plant Cell* 19, 3058 LP – 3079.
- Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., et al. (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* 5, 3978.
- Jahrman, T., Bastida, M., Pineda, M., Gasol, E., Ludevid, M.D., Palacín, M., and Puigdomènech, P. (2005). Studies on the function of TM20, a transmembrane protein present in cereal embryos. *Planta* 222, 80–90.
- Ju, C., Van De Poel, B., Cooper, E.D., Thierer, J.H., Gibbons, T.R., Delwiche, C.F., and Chang, C. (2015). Conservation of ethylene as a plant hormone over 450 million years of evolution. *Nat. Plants* 1, 1–7.
- Kato, H., Ishizaki, K., Kouno, M., Shirakawa, M., Bowman, J.L., Nishihama, R., and 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 Genet.* 11, 1–26.
- Kato, H., Nishihama, R., Weijers, D., and Kohchi, T. (2018). Evolution of nuclear auxin signaling: Lessons from genetic studies with basal land plants. *J. Exp. Bot.* 69, 291–301.

- Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J., and Gray, M.W. (2005). The tree of eukaryotes. *Trends Ecol. Evol.* 20, 670–676.
- Kenrick, P., and Crane, P.R. (1997). The origin and early evolution of plants on land. *Nature* 389, 33–39.
- Koumandou, V.L., Wickstead, B., Ginger, M.L., van der Giezen, M., Dacks, J.B., and Field, M.C. (2013). Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit. Rev. Biochem. Mol. Biol.* 48, 373–396.
- Lavy, M., Prigge, M.J., Tao, S., Shain, S., Kuo, A., Kirchsteiger, K., and Estelle, M. (2016). Constitutive auxin response in *Physcomitrella* reveals complex interactions between Aux/IAA and ARF proteins. *Elife* 5, 1–22.
- Liang, Z., Geng, Y., Ji, C., Du, H., Wong, C.E., Zhang, Q., Zhang, Y., Zhang, P., Riaz, A., Chachar, S., et al. (2019). *Mesostigma viride* Genome and Transcriptome Provide Insights into the Origin and Evolution of Streptophyta. *Adv. Sci.* 201901850
- Ljung, K. (2013). Auxin metabolism and homeostasis during plant development. *Development* 140, 943–950.
- Ludwig-Müller, J. (2011). Auxin conjugates: their role for plant development and in the evolution of land plants. *J. Exp. Bot.* 62, 1757–1773.
- Lughadha, E.N., Govaerts, R., Belyaeva, I., Black, N., Lindon, H., Allkin, R., Magill, R.E., And Nicolson, N. (2016). Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. *Phytotaxa* 272, 82.
- Luschnig, C., and Vert, G. (2014). The dynamics of plant plasma membrane proteins: PINs and beyond. *Development* 141, 2924–2938.
- Marin, B. (2012). Nested in the Chlorellales or Independent Class? Phylogeny and Classification of the Pedinophyceae (Viridiplantae) Revealed by Molecular Phylogenetic Analyses of Complete Nuclear and Plastid-encoded rRNA Operations. *Protist* 163, 778–805.
- Mashiguchi, K., Tanaka, K., Sakai, T., Sugawara, S., Kawaide, H., Natsume, M., Hanada, A., Yaeno, T., Shirasu, K., Yao, H., et al. (2011). The main auxin biosynthesis pathway in *Arabidopsis*. *Proc. Natl. Acad. Sci.* 108, 18512–18517.
- Menand, B., Yi, K., Jouannic, S., Hoffmann, L., Ryan, E., Linstead, P., Schaefer, D.G., and Dolan, L. (2007). An Ancient Mechanism Controls the Development of Cells with a Rooting Function in Land Plants. *Science* 316, 1477 – 1480.
- Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z., Schneider, H., and Donoghue, P.C.J. (2018). The timescale of early land plant evolution. *Proc. Natl. Acad. Sci.* 201719588.
- Mravec, J., Kubeš, M., Bielach, A., Gaykova, V., Petrášek, J., Skúpa, P., Chand, S., Benková, E., Zažímalová, E., and Friml, J. (2008). Interaction of PIN and PGP transport mechanisms in auxin distribution-dependent development. *Development* 135, 3345 LP – 3354.
- Nishiyama, T., Sakayama, H., de Vries, J., Buschmann, H., Saint-Marcoux, D., Ullrich, K.K., Haas, F.B., Vanderstraeten, L., Becker, D., Lang, D., et al. (2018). The *Chara* Genome: Secondary Complexity and Implications for Plant Terrestrialization. *Cell* 174, 448–464.e24.
- Nobuta, K., Okrent, R.A., Stoutemyer, M., Rodibaugh, N., Kempema, L., Wildermuth, M.C., and Innes, R.W. (2007). The GH3 Acyl Adenylase Family Member PBS3 Regulates Salicylic Acid-Dependent Defense Responses in *Arabidopsis*. *Plant Physiol.* 144, 1144–1156.
- Noh, B., Murphy, A.S., and Spalding, E.P. (2001). Multidrug Resistance-Like Genes of *Arabidopsis* Required for Auxin Transport and Auxin-Mediated Development. *Plant Cell* 13, 2441.
- Normanly, J., Grisafi, P., Fink, G.R., and Bartel, B. (1997). *Arabidopsis* Mutants Resistant to the Auxin Effects of Indole-3-Acetonitrile Are Defective in the Nitrilase Encoded by the NIT1 Gene. *Plant Cell* 9, 1781.
- Ohtaka, K., Hori, K., Kanno, Y., Seo, M., and Ohta, H. (2017). Primitive Auxin Response without TIR1 and Aux/IAA in the Charophyte Alga *Klebsormidium nitens*. *Plant Physiol.* 174, 1621–1632.
- Okada, K., Ueda, J., Komaki, M.K., Bell, C.J., and Shimura, Y. (1991). Requirement of the Auxin Polar Transport System in Early Stages of *Arabidopsis* Floral Bud Formation. *Plant Cell* 3, 677.
- Okrent, R.A., Brooks, M.D., and Wildermuth, M.C. (2009). *Arabidopsis* GH3.12 (PBS3) Conjugates Amino Acids to 4-Substituted Benzoates and Is Inhibited by Salicylate. *J. Biol. Chem.* 284, 9742–9754.
- One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685.
- Paponov, I.A., Dindas, J., Król, E., Friz, T., Budnyk, V., Teale, W., Paponov, M., Hedrich, R., and Palme, K. (2019). Auxin-Induced Plasma Membrane Depolarization Is Regulated by Auxin Transport and Not by AUXIN BINDING PROTEIN1. *Front. Plant Sci.* 9.
- Parry, G., Calderon-Villalobos, L.I., Prigge, M., Peret, B., Dharmasiri, S., Itoh, H., Lechner, E., Gray, W.M., Bennett, M., and Estelle, M. (2009). Complex regulation of the TIR1/AFB family of auxin receptors. *Proc. Natl. Acad. Sci.* 106, 22540–22545.
- Prigge, M.J., Lavy, M., Ashton, N.W., and Estelle, M. (2010). *Physcomitrella* patens auxin-resistant mutants affect conserved elements of an auxin-signaling pathway. *Curr. Biol.* 20, 1907–1912.
- Proust, H., Hoffmann, B., Xie, X., Yoneyama, K., Schaefer, D.G., Yoneyama, K., Nogué, F., and Rameau, C. (2011).

- Strigolactones regulate protonema branching and act as a quorum sensing-like signal in the moss *Physcomitrella patens*. *Development* 138, 1531 LP – 1539.
- Proust, H., Honkanen, S., Jones, V.A.S., Morieri, G., Prescott, H., Kelly, S., Ishizaki, K., Kohchi, T., and Dolan, L. (2016). RSL Class I Genes Controlled the Development of Epidermal Structures in the Common Ancestor of Land Plants. *Curr. Biol.* 26, 93–99.
- Puttick, M.N., Morris, J.L., Williams, T.A., Cox, C.J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Schneider, H., Pisani, D., et al. (2018). The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Curr. Biol.* 28, 733–745.
- Robert, H.S., Grunewald, W., Sauer, M., Cannoot, B., Soriano, M., Swarup, R., Weijers, D., Bennett, M., Boutilier, K., and Friml, J. (2015). Plant embryogenesis requires AUX/LAX-mediated auxin influx. *Development* 142, 702–711.
- Robert, S., Kleine-Vehn, J., Barbez, E., Sauer, M., Paciorek, T., Baster, P., Vanneste, S., Zhang, J., Simon, S., Čovanová, M., et al. (2010). ABP1 Mediates Auxin Inhibition of Clathrin-Dependent Endocytosis in *Arabidopsis*. *Cell* 143, 111–121.
- Rubery, P.H., and Sheldrake, A.R. (1974). Carrier-mediated auxin transport. *Planta*.
- Ruggiero, M.A., Gordon, D.P., Orrell, T.M., Bailly, N., Bourgoin, T., Brusca, R.C., Cavalier-Smith, T., Guiry, M.D., and Kirk, P.M. (2015). A higher level classification of all living organisms. *PLoS One* 10, 1–60.
- Santner, A., and Estelle, M. (2009). Recent advances and emerging trends in plant hormone signalling. *Nature* 459, 1071–1078.
- Sauer, M., and Kleine-Vehn, J. (2011). AUXIN BINDING PROTEIN1: The outsider. *Plant Cell* 23, 2033–2043.
- Seo, M., Akaba, S., Oritani, T., Delarue, M., Bellini, C., Caboche, M., and Koshiba, T. (1998). Higher Activity of an Aldehyde Oxidase in the Auxin-Overproducing superroot1 Mutant of *Arabidopsis thaliana*. *Plant Physiol.* 116, 687–693.
- Simpson, A.G.B., and Roger, A.J. (2004). The real ‘kingdoms’ of eukaryotes. *Curr. Biol.* 14, R693–R696.
- Skokan, R., Medvecká, E., Viaene, T., Vosolsobě, S., Zwiewka, M., Müller, K., Skůpa, P., Karady, M., Zhang, Y., Janacek, D.P., et al. (2019). PIN-driven auxin transport emerged early in streptophyte evolution. *Nat. Plants* 5, 1114–1119.
- Staswick, P.E., Tiryaki, I., and Rowe, M.L. (2002). Jasmonate Response Locus JAR1 and Several Related *Arabidopsis* Genes Encode Enzymes of the Firefly Luciferase Superfamily That Show Activity on Jasmonic, Salicylic, and Indole-3-Acetic Acids in an Assay for Adenylation. *Plant Cell* 14, 1405–1415.
- Staswick, P.E., Serban, B., Rowe, M., Tiryaki, I., Maldonado, M.T., Maldonado, M.C., and Suza, W. (2005). Characterization of an *Arabidopsis* Enzyme Family That Conjugates Amino Acids to Indole-3-Acetic Acid. *Plant Cell* 17, 616–627.
- Stepanova, A.N., Robertson-Hoyt, J., Yun, J., Benavente, L.M., Xie, D.-Y., Doležal, K., Schlereth, A., Jürgens, G., and Alonso, J.M. (2008). TAA1-Mediated Auxin Biosynthesis Is Essential for Hormone Crosstalk and Plant Development. *Cell* 133, 177–191.
- Sun, Y., Harpazi, B., Wijerathna-Yapa, A., Merilo, E., de Vries, J., Michaeli, D., Gal, M., Cuming, A.C., Kollist, H., and Mosquna, A. (2019). A ligand-independent origin of abscisic acid perception. *Proc. Natl. Acad. Sci.* 116, 24892–24899.
- Swarup, R., and Péret, B. (2012). AUX/LAX family of auxin influx carriers—an overview. *Front. Plant Sci.* 3.
- Timme, R.E., and Delwiche, C.F. (2010). Uncovering the evolutionary origin of plant molecular processes: comparison of Coleochaete (Coleochaetales) and Spirogyra (Zygnematales) transcriptomes. *BMC Plant Biol.* 10, 96.
- Tiwari, S.B., Hagen, G., and Guilfoyle, T. (2003). The Roles of Auxin Response Factor Domains in Auxin-Responsive Transcription. *Plant Cell* 15, 533–543.
- Tiwari, S.B., Hagen, G., and Guilfoyle, T.J. (2004). Aux/IAA Proteins Contain a Potent Transcriptional Repression Domain. *Plant Cell* 16, 533–543.
- Ugartechea-Chirino, Y., Swarup, R., Swarup, K., Peret, B., Whitworth, M., Bennett, M., and Bougourd, S. (2010). The AUX1 LAX family of auxin influx carriers is required for the establishment of embryonic root cell organization in *Arabidopsis thaliana*. *Ann. Bot.* 105, 277–289.
- Ulmasov, T., Hagen, G., and Guilfoyle, T.J. (1999). Activation and repression of transcription by auxin-response factors. *Proc. Natl. Acad. Sci. U. S. A.* 96, 5844–5849.
- Verrier, P.J., Bird, D., Burla, B., Dassa, E., Forestier, C., Geisler, M., Klein, M., Kolukisaoglu, Ü., Lee, Y., Martinoia, E., et al. (2008). Plant ABC proteins - a unified nomenclature and updated inventory. *Trends Plant Sci.* 13, 151–159.
- Viaene, T., Delwiche, C.F., Rensing, S.A., and Friml, J. (2013). Origin and evolution of PIN auxin transporters in the green lineage. *Trends Plant Sci.* 18, 5–10.
- Weijers, D., and Wagner, D. (2016). Transcriptional Responses to the Auxin Hormone. *Annu. Rev. Plant Biol.* 67, 539–574.
- Weijers, D., Sauer, M., Meurette, O., Friml, J., Ljung, K., Sandberg, G., Hooykaas, P., and Offringa, R. (2005). Maintenance of Embryonic Auxin Distribution for Apical-Basal Patterning by PIN-FORMED–Dependent Auxin Transport

- in *Arabidopsis*. *Plant Cell* 17, 2517–2526.
- Weijers, D., Schlereth, A., Ehrismann, J.S., Schwank, G., Kientz, M., and Jürgens, G. (2006). Auxin Triggers Transient Local Signaling for Cell Specification in *Arabidopsis* Embryogenesis. *Dev. Cell* 10, 265–270.
- Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. U. S. A.* 111, E4859–E4868.
- Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci.* 74, 5088–5090.
- Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* 87, 4576–4579.
- Woo, E.-J. (2002). Crystal structure of auxin-binding protein 1 in complex with auxin. *EMBO J.* 21, 2877–2885.
- Žárský, V., Cvrčková, F., Potocký, M., and Hála, M. (2009). Exocytosis and cell polarity in plants - exocyst and recycling domains. *New Phytol.* 183, 255–272.
- Zhang, J., Lin, J.E., Harris, C., Campos Mastrotti Pereira, F., Wu, F., Blakeslee, J.J., and Peer, W.A. (2016). DAO1 catalyzes temporal and tissue-specific oxidative inactivation of auxin in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* 113, 11010 LP – 11015.
- Zhang, L., Zhang, F., Melotto, M., Yao, J., and He, S.Y. (2017). Jasmonate signaling and manipulation by pathogens and insects. *J. Exp. Bot.* erw478.
- Zhao, Y. (2002). Trp-dependent auxin biosynthesis in *Arabidopsis*: involvement of cytochrome P450s CYP79B2 and CYP79B3. *Genes Dev.* 16, 3100–3112.
- Zhao, Y. (2012). Auxin Biosynthesis: A Simple Two-Step Pathway Converts Tryptophan to Indole-3-Acetic Acid in Plants. *Mol. Plant* 5, 334–338.
- Zhao, Y. (2014). Auxin Biosynthesis. *Arab. B.* 12, e0173.



Chapter 2

High-resolution and deep phylogenetic reconstruction of ancestral states from large transcriptomic data sets

Sumanth Kumar Mutte and Dolf Weijers

Laboratory of Biochemistry, Wageningen University, Wageningen, the Netherlands

Modified version of this chapter has been published as:

Mutte, S. K., & Weijers, D. (2020). High-resolution and deep phylogenetic reconstruction of ancestral states from large transcriptomic data sets. *Bio-Protocol*, in press.



Phylogenetics is an important area of evolutionary biology that helps to understand the origin and divergence of genes, genomes and species. Building meaningful phylogenetic trees is needed for the accurate reconstruction of the past. To achieve a correct phylogenetic understanding of genes or proteins, reliable and robust methods are needed to construct meaningful trees. With the rapidly increasing genome and transcriptome sequencing data, there is a need for efficient and accurate methodologies for ancestral state reconstruction. Currently available methods are mostly specific for certain gene families, and require substantial adaptation for their application to other gene families. Hence, a generalized framework is essential to utilize large transcriptome resources such as OneKP and MMETSP. Here, we have developed a flexible yet efficient method, based on core strengths such as emphasis on being inclusive in homolog selection, and defining orthologs based on multi-layered inferences. We illustrate how specific steps can be modified to fit the needs of any protein family under consideration. We also demonstrate the success of this protocol by studying and testing the orthologs of multiple gene families. Taken together, we present a protocol for reconstructing the ancestral states of various domains and proteins across multiple kingdoms of eukaryotes, using thousands of transcriptomes.

Introduction


Phylogenetic trees are fundamental to understand the evolution of genes, gene families, species, phyla and even kingdoms. They help us depict the diversity and also resolve the differences at various levels. For example, at protein level, they help us identify orthologous groups based on amino acid differences across various species. Earlier, phylogenetic trees were constructed based on few gene/protein sequences from a limited number of species. With the ever-growing sequencing data, as more and more genomes and transcriptomes are becoming accessible, there is tremendous potential, for e.g. discovery of new lineages, ‘gap-filling’ in phylogenies and hence, an improved understanding of biology (Burki et al., 2019; Levy and Myers, 2016).

In the last decade, many efforts have been made towards defining transcriptomes of hundreds (or even thousands) of species due to the popularity of RNA-Seq (Stark et al., 2019). Transcriptomes provide a quick insight into the (expressed) gene content of a genome. Even though the individual transcriptomes do not cover the entire gene content of an organism, combining them from multiple cells, tissues and conditions, may comprise the majority of the transcribed genes of that species. Hence, it is a relatively straightforward approach to sequence and assemble a transcriptome without *a priori* knowledge of the genome. The current-day long-read and single-cell RNA-sequencing technologies make it even easier to build a complete transcriptome (Wang et al., 2016). Utilizing these technological advances, two large transcriptome sequencing projects, 1000 plant transcriptomes (OneKP; Carpenter et al., 2019; Matasci et al., 2014; One Thousand Plant Transcriptomes Initiative, 2019) and Marine Micro Eukaryote Transcriptome Sequencing Project (MMETSP; Keeling et al., 2014), were developed. OneKP represents majority of the Archaeplastida, whereas MMETSP covers majority of the SAR (Stramenopila, Alveolata and Rhizaria) group and other (unidentified) phyla in Chromista.

From their inception, diverse approaches have been developed and applied to these transcriptomes and estimate the ancestral states of various genes across multiple classes, families and even phyla (Li et al., 2014; Wickett et al., 2014; Yerramsetty et al., 2016). The majority of these methods focus on one gene family, and need substantial modifications in methodology to apply them to other gene families. Moreover, the methods used are neither inclusive nor robust in terms of multi-layered inferences. The orthologous inferences are based on only one evidence, either Best Bi-directional Hit (BBH) or protein domains or simple phylogenies based on few species. To overcome these disadvantages, we developed a unified framework to build high-resolution phylogenies that utilize the rich OneKP and MMETSP transcriptome resources. This new method is not only inclusive, but also utilizes multi-layered orthology to interpret phylogenies with high confidence, leading to the identification of new (sub)classes of orthologs.

Overview of the protocol

The current protocol is developed to reconstruct ancestral states and high-resolution phylogenetic trees of various gene families using transcriptomes and/or proteomes. Ancestral state represents the minimal gene complement at each evolutionary node, where species-specific gene duplications



and (or) losses would have modified the gene complement in individual species. Hence, selecting the correct, orthologous as well as diverse, sequences is a crucial step in such a deep phylogenetic tree construction. This protocol is built on three core strengths: (1) Inclusive: Include more sequences at the start with liberal parameters, and remove sequences as one goes through various steps in the pipeline, resulting in a high-quality logical sequence set for phylogenetic tree construction. (2) Multi-layered: Multiple levels of orthology confirmation, i.e. based on the domain architecture, reciprocal BLAST and the phylogenetic tree. (3) Robust: No limitations on length of the protein or the number of sequences used in the phylogeny, with suggestions on alternate analysis packages tested in various steps. Overall, the protocol comprises 14 steps that are divided into three sections: Homolog identification (Steps 1-5), Ortholog detection (Steps 6-8) and Phylogeny construction (Steps 9-14). All the general parameters and recommendations for the respective steps are indicated below. Gene family specific parameters are mentioned in the corresponding chapters (Chapters 3, 4 and 5).

Data and software used

DATA

- OneKP dataset (1000 plant transcriptomes project): Contains 1341 transcriptomes from 1179 species covering all the major classes of land plants, green algae, red algae and glaucophytes (Carpenter et al., 2019; Matasci et al., 2014; One Thousand Plant Transcriptomes Initiative, 2019); http://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/oneKP_capstone_2019
- MMETSP dataset (Marine Microbial Eukaryote Transcriptome Sequencing Project): Contains 678 transcriptomes from 410 species covering all the major classes of Stramenopila and Alveolata (SAR group) and many unclassified (unicellular) marine eukaryotes (Keeling et al., 2014); <https://gold.jgi.doe.gov/study?id=G0128947>

SOFTWARE

- tblastn and blastp from BLAST+ module v2.9.0 (Camacho et al., 2009) (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>)
- faSomeRecords: Linux binary from UCSC (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/)
- TransDecoder v5.5.0 (Haas et al., 2013) ([transdecoder.github.io](https://github.com/transdecoder/transdecoder))
- MEME motif discovery v5.1.0 (Bailey et al., 2009) (<http://meme-suite.org/>)
- ScanProsite web-tool (<https://prosite.expasy.org/scanprosite>)
- InterProScan v5.38-76.0 (Jones et al., 2014) (<https://github.com/ebi-pf-team/interproscan>)
- MAFFT v7 (Katoh and Standley, 2013) (<https://mafft.cbrc.jp/alignment/software/>)
- JalView (Waterhouse et al., 2009) (<https://www.jalview.org/>)
- ModelFinder (Kalyaanamoorthy et al., 2017) (accessed as in-built module from IQ-TREE)

- ModelTest-NG (Darriba et al., 2019) (<https://github.com/ddarriba/modeltest>)
- PartitionFinder v2 (Lanfear et al., 2012) (<http://www.robertlanfear.com/partitionfinder/>)
- IQ-TREE v1.6.12 (Nguyen et al., 2015) (<http://www.iqtree.org>)
- RAxML v8 (Stamatakis, 2014) (<https://cme.h-its.org/exelixis/web/software/raxml/index.html>)
- PhyML v3.3 (Guindon et al., 2010) (<https://github.com/stephaneguindon/phyml>)
- MrBayes v3.2.7 (Ronquist et al., 2012) (<https://github.com/NBISweden/MrBayes>)
- iTOL v4 (Letunic and Bork, 2019) (<https://itol.embl.de>)
- Various scripts used for automating certain steps in the protocol are available through GitHub (<https://github.com/sumanthmutte/Phylogenomics>)

Procedure

Commands used, along with the parameters used in each step of the protocol, with step numbers corresponding to Figure 1 are given below. Before starting the protocol, we first created a BLAST database for each transcriptome or proteome. This was carried out only once for each transcriptome or proteome using the *makeblastdb* function, where '-in' takes a FASTA file of the transcriptome, or the proteome and '-dbtype' is the database type with 'nucl' and 'prot' for transcriptomes and proteomes, respectively.

```
$ makeblastdb -dbtype nucl -in transcriptome.fasta
$ makeblastdb -dbtype prot -in proteome.fasta
```

HOMOLOG IDENTIFICATION

1. To perform a BLAST search to the respective database(s), we created a query protein sequence file (in FASTA format), with sequences from (relatively) well-annotated genomes and from a diverse range of species, if present, across multiple kingdoms. A list of various species used along with a link to the sequence data resource is available in supplementary information.

2. We have used this query sequence file (-query) to perform the BLAST search using *tblastn* and *blastp* modules, against transcriptome and proteome databases (-db), respectively. The E-value cut-off (-evalue) was less than 0.01 and the output (-out) was saved in a tab-delimited text file indicated with '-outfmt 6' followed by the columns to be saved in that file. The remainder of the parameters were kept at default settings.

```
$ tblastn -query filename.fa -db transcriptome.fasta -out output.blast -evalue 0.01
-outfmt '6 qseqid sseqid slen qstart qend sstart send eval bitscore score length pident nident
positive ppos mismatch gaps frames qcovs qcovhsp sseq'
$ blastp -query filename.fa -db proteome.fasta -out output.blast -evalue 0.01 -outfmt '6
qseqid sseqid slen qstart qend sstart send eval bitscore score length pident nident positive ppos
mismatch gaps frames qcovs qcovhsp sseq'
```

3. The BLAST output contains all the scoring information about the subject (transcript/protein) sequence that has a similarity to the corresponding query sequence. To retrieve the subject

sequence identifiers from the BLAST output, we have used the ‘*cut*’, ‘*sort*’ and ‘*uniq*’ functions of Linux BASH shell (terminal). ‘*cut*’ takes the BLAST output (output.blast) from the previous step, and takes the second column (-f2) i.e. subject sequence identifiers and sends/pipes them (|) to the ‘*sort*’ function. After sorting, they are passed on to the ‘*uniq*’ function to remove the duplicates and the output is written to the file (SubjectIdentifiers.txt).

```
$ cut -f2 output.blast | sort | uniq > SubjectIdentifiers.txt
```

4. We further used these identifiers (SubjectIdentifiers.txt) to extract the corresponding transcript (SelectedTranscripts.fasta) or protein sequences (SelectedProteins.fasta) from the respective transcriptome or proteome by running the ‘*faSomeRecords*’ program.

```
$ faSomeRecords transcriptome.fasta SubjectIdentifiers.txt SelectedTranscripts.fasta
$ faSomeRecords proteome.fasta SubjectIdentifiers.txt SelectedProteins.fasta
```

5. Since protein sequences are more informative due to the higher number of site patterns, we decided to use protein sequences for phylogeny construction. Hence, the protein sequences from the previous step were directly used for further analysis. Whereas, the transcript sequences were first translated to protein sequences using the program *TransDecoder* with default settings. In the situations where the protein sequences resulted in poor bootstrap support on trees, we have also generated trees with CDS (DNA) sequences, which were also obtained from the *TransDecoder* program. In the first step (TransDecoder.LongOrfs), the longest Open Reading Frames (ORFs) of at least 100 amino acids in length) of the transcript are determined. In the second step (TransDecoder.Predict), the CDS and the corresponding amino acid sequences of these ORFs were determined.

```
$ perl TransDecoder.LongOrfs -t SelectedTranscripts.fasta
$ perl TransDecoder.Predict -t SelectedTranscripts.fasta
```

ORTHOLOG DETECTION

6. Not all the sequences that have an E-value < 0.01 are true orthologs of a query protein. Hence, we need additional filters to remove non-orthologs. One such filter is the presence of the same domains in orthologous proteins. For some well-annotated proteins (e.g. Auxin Response Factors, Kinases etc.), domain information is readily available in the *InterPro* domain database. Hence, the protein sequences from the previous step (-i SelectedProteins.fasta) were scanned for the presence of known domains using *InterProScan* tool (interproscan.sh), which produces a tab-delimited (TSV) file as well as HTML/XML files (-f TSV,HTML,XML), with all the domains identified along with the corresponding InterPro identifiers (-iprlookup) in each protein sequence. A Python script was developed (InterproscanSummary.py; <https://github.com/sumanthmutte/Phylogenomics>) to process this TSV file, in order to extract the final set of protein sequences that have the domains of interest. We have used this approach for the majority of the proteins studied in this thesis. *InterProScan* is a time-consuming step, hence we used pre-annotated data where available, or reduced the number of databases to scan (using -appl Pfam,CDD setting), in order

to save time. In some cases, we split the data in smaller batches and ran on multiple processors.

```
$ interproscan.sh -f TSV,HTML,XML -iplookup -i SelectedProteins.fasta
$ python InterproscanSummary.py
```

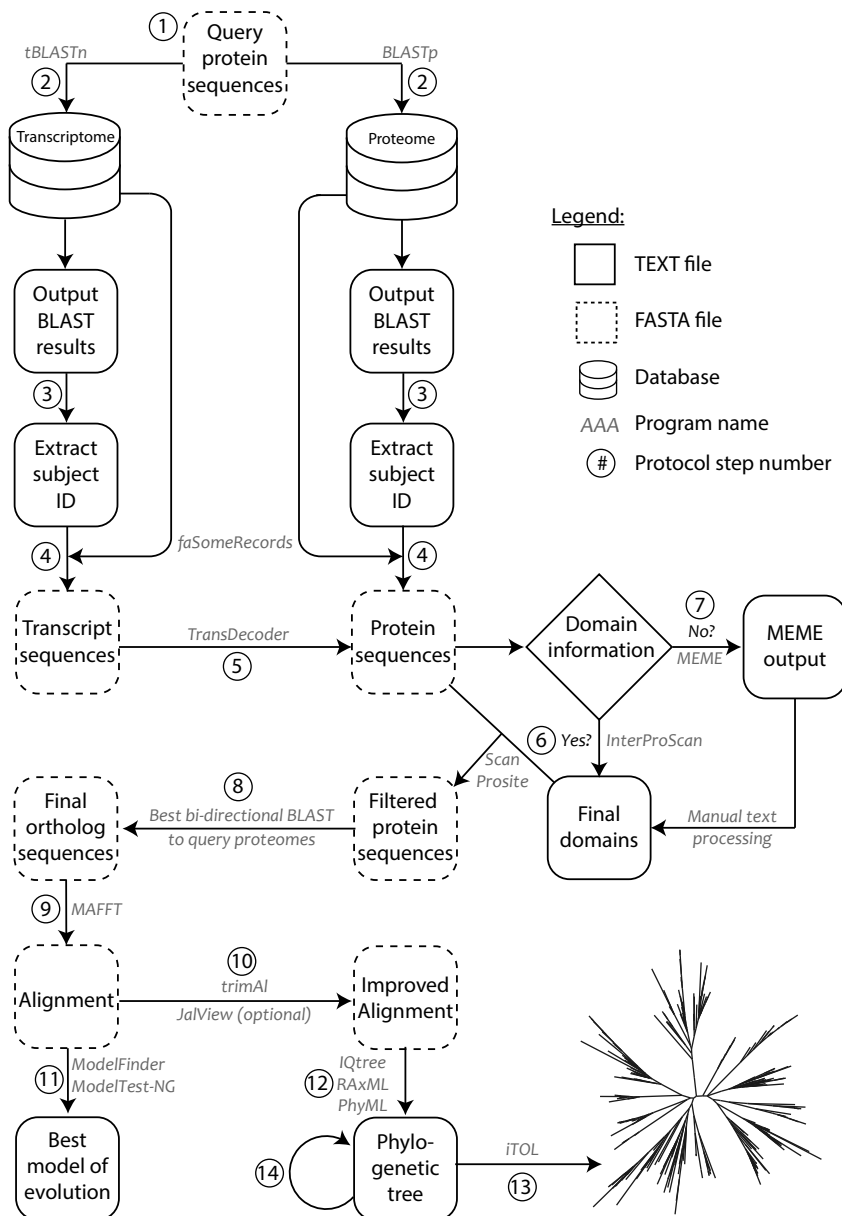


Figure 1: Methodology schematic showing various steps of the protocol used for ortholog identification and phylogenetic tree construction. Circled numbers correspond to the various steps of the protocol as indicated in the procedure. Programs or software used are indicated next to the arrows in grey. File formats for text and FASTA are depicted as shown in legend.

7. Certain proteins (e.g. SOSEKI in *Arabidopsis*; Yoshida et al., 2019) lack annotated (functional) domain information. To predict the conserved motifs/domains in those proteins, we have used the *MEME* program, with Zero or One Occurrence Per Sequence criteria (-mod zoops) and a minimum width of 10 (-minw), with a maximum of 10 motifs predicted per set (-nmotifs). *MEME* outputs the motifs along with their patterns in HTML/TEXT format. We then used these motif patterns in *ScanProsite* web-tool to identify the domains in the protein sequences that do not have annotated domains. We have applied this approach successfully to annotate the SOSEKI protein family and identify its orthologs (van Dop et al., 2020; Chapter 3).

```
$ meme ProteinSeq.fa -o OutputName-protein -mod zoops -nmotifs 10 -minw 10
```

8. After selecting the protein sequences that have the domains of interest, they were queried back to the proteomes of the species used in step-1 to confirm the orthologous relationships using best Bi-directional BLAST Hits (BBH) strategy. Here we have used the option of maximum target sequences or the number of best hits in the output (-max_target_seqs) set to 1, or sometimes 2 when domains are abundant in the genome (for e.g. bHLH), with E-value < 0.01 (-evalue). This final set of proteins that have hits with the protein under consideration were regarded as the 'true' orthologous proteins for further analysis. Output is recorded in a TSV file, same as in step-2 (-outfmt 6).

```
$ blastp -query filename.fa -db ArabidopsisProteome.fasta -out BBhits.blastp -max_target_seqs 1 -evalue 0.01 -outfmt '6 qseqid sseqid slen qstart qend sstart send eval evalue bitscore score length pident nident positive ppos mismatch gaps frames qcovs qcovhsp sseq'
```

PHYLOGENY CONSTRUCTION

9. These 'true' sets of orthologs were used for alignment followed by the phylogenetic tree construction. *MAFFT* was used to align protein sequences. The *E-INS-i* (--genafpair) algorithm was used while aligning proteins with multiple domains separated by poorly conserved sequences (e.g. ARFs, Aux/IAAs; Chapter 4), whereas *G-INS-i* (--globalpair) was used while aligning only domain-specific sequences (e.g. PB1 domain; Chapter 5). An iterative refinement method was used in both cases, with a maximum of 1000 iterations (--maxiterate 1000), after which the final alignment is written to a FASTA file (output_file).

```
$ mafft --genafpair --maxiterate 1000 input_file > output_file
```

```
$ mafft --globalpair --maxiterate 1000 input_file > output_file
```

10. Once the alignments were generated, we used *trimAl* to remove the sequence positions (columns) with more than 50%-80% gaps, as they are considered to lack phylogenetic signal. Hence, for phylogenetic tree construction, we have used only the sequence positions without spurious gaps. There are various tools specialized for the clean-up of the alignment, such as *GBLOCKS* (Talavera and Castresana, 2007) and *Guidance* (Sela et al., 2015). However, a simple gap-based trimming in *trimAl* resulted in (almost) the same quality of alignment and tree topology when compared to these specialized tools. Hence, we used *trimAl* for alignment clean-

up throughout this study. A gap-threshold of 0.2 (-gt 0.2), is set to remove all positions in the alignment with gaps in 80% (or more) of the sequences. For the gene families that have moderately conserved domains (e.g. ARF, Aux/IAA), we have used a threshold of 0.3 or 0.4, whereas for poorly conserved domains (e.g. PB1) it was set at 0.2, and for highly conserved proteins (e.g. ROP, ROPGEF) it was set between 0.6 to 0.8. An additional (optional) check is kept in place, where the sequences that were shorter than 1/4th of the average sequence length were further removed in *JalView*.

```
$ trimal -in inputfile.fa -out outputfile.fa -fasta -gt 0.2
```

11. We then used this ‘clean’ alignment to identify the most appropriate model of evolution for each protein family. *ModelFinder* and *ModelTest-NG* were used to predict the best model based on the Akaike- and Bayesian- Information Criterion (AIC and BIC). For majority of the protein families, both programs provided the same models as the best models. The situations where there was a mis-match between the two programs, we have used a third program (either PartitionFinder or a Perl script from *RAxML* distribution) to decide on the best model based on the majority rule. As expected, various proteins evolved differently, leading to different models of evolution. Models used for the phylogeny construction of respective protein families are discussed later in the corresponding chapters. *ModelFinder* was run as a part of IQ-TREE, hence it did not require any additional steps. *ModelTest-NG* required the type (either amino acid or nucleotide -d) of input dataset (-i INFILE) and writes the statistics and the best model to the output file (-o OUTFILE). *PartitionFinder* requires the alignment, in the PHYLIP format (instead of FASTA format as in others), placed in the folder ‘partition_finder_models’, where the output statistics and best model were also recorded. FASTA to PHYLIP format conversion was made through the Perl script (fasta2relaxedPhylip.pl), which takes input FASTA (-f input.fa) and writes the output in PHYLIP format (-o output.phylip).

```
$ modeltest-ng -d aa -i INFILE -o OUTFILE
$ perl RAXML_ProteinModelSelection.pl alignment.fasta
$ perl fasta2relaxedPhylip.pl -f input.fa -o output.phylip
$ python PartitionFinderProtein.py partition_finder_models
```

12. Phylogenetic trees were built mainly using *IQ-TREE* and *RAxML* based on the ‘clean’ alignment produced in step-10 and the evolutionary model predicted in step-11. For the phylogenetic trees made through *IQ-TREE*, we have used 1000 rapid bootstraps (-bb 1000) and SH-like approximate Likelihood Ratio Test (-aLRT 1000), combined with automatic model finding through *ModelFinder* (-m MFP+MERGE). For the trees made with *RAxML*, we have also used rapid bootstrapping and Maximum Likelihood search in the same run (-f a) but with an extended majority rule (-# autoMRE) based bootstopping criteria. In addition, we gave a random seed number (-x and -p) to turn-on rapid bootstrapping and parsimony inference, whereas -m takes in the model from the previous step. For trees with very poor bootstrap support for majority of the branches, we used another phylogenetic tree construction program, *PhyML*, with 100 bootstrap replicates (-b 100), empirical amino-acid frequencies (-f e), gamma shape

parameter estimated from maximum likelihood (-a e) and the topology was searched based on the sub-tree pruning and re-grafting approach (-s SPR). After running these multiple programs, the trees obtained were compared to understand the overall topology based on the congruent branches. We have also tried and tested various Bayesian approaches (using *MrBayes*), but the trees never converged even after months of computation, and provided various incongruent topologies. Hence, all the analyses in this thesis were performed with Maximum Likelihood approaches.

```
$ iqtree -s CleanAlignment.fa -pre OutputName -alrt 1000 -bb 1000 -m MFP+MERGE
$ raxmlHPC-PTHREADS-AVX2 -f a -x 12345 -p 12345 -j -# autoMRE -m
PROTGAMMAJTT -s CleanAlignment.fa -n OutputName
$ PhyML-3.1_linux64 -i CleanAlignment.fa -d aa -b 100 -m JTT -f e -s SPR -a e
```

13. All the final phylogenetic trees mentioned in various chapters were visualized using the iTOL webserver. Various datasets were visualized on the phylogenetic trees. Protein domain information from the *InterProScan* or *MEME*, sequence length from *TransDecoder* and clade/taxonomy information from OneKP and MMETSP databases were generated following the instructions provided in the iTOL documentation.

14. Once the trees were obtained, they were manually checked for errors. Branches with long branch attraction, or partial sequences or any misplaced taxa were manually removed. If the proportion of these misplaced branches was too high, we re-analysed the phylogeny with more sequences from other species, as well as by removing the spurious sequences. These steps were repeated until we obtained better trees that were not only supported by good bootstraps but also obeys the taxonomy of those phyla.

Limitations and Conclusions

Due to the generalized nature of the method, it was difficult to automate the complete protocol. Hence, wherever possible, the method was simplified with scripts/commands dedicated for fast and parallel processing. On the other hand, it gave control over the decision-making process based on the protein under consideration. When dealing with highly redundant protein families, we removed highly similar proteins (>90% similarity), prior to phylogeny, which reduced the (computational) time without losing accuracy. In many cases we observed that the best-hit in *reciprocal-BLAST* is not really a BBH, as sometimes a second hit was still the best one due to one or few amino acid difference(s) (especially in proteins with common domains e.g. bHLH or PB1). Hence, in those cases we considered two best hits and used both for phylogeny construction. The false positive orthologs were eventually placed in the outgroup (or at least separate from the ingroup) in the phylogenetic tree. As we were dealing with transcriptomes, we could not predict the actual gene copy number in each species, but only the ancestral copy number for that class or phylum, by comparing the ancestral copies across the majority of the species in that phylum. Another issue of dealing with (low-depth) transcriptomes was that we found many partial transcripts leading to the truncated proteins/domains, or we might fail to

identify the transcripts that were not expressed in that particular tissue or condition. In that regard, combining ortholog sequence information from multiple transcriptomes or species of various families is mandatory to confirm the ancestral state for each class or phylum.

Based on this protocol and the guidelines mentioned above, we have reconstructed the ancestral states of various protein families along with their orthologs in a ‘deep’ phylogenetic space, across multiple kingdoms. In the consecutive chapters, we demonstrate how this method was implemented for proteins that are well-defined with known domains, novel proteins with unknown domains, poorly conserved domains and phylum/kingdom-specific proteins that (dis)appeared at various stages in evolution. This approach was successfully applied for the core proteins of the auxin signalling (Nuclear Auxin Pathway (NAP)) and biosynthesis pathways. NAP includes Auxin Response Factor (ARF), Auxin/Indole-3-Acetic-Acid (Aux/IAA) and Transport Inhibitor Response 1/Auxin-signalling F-Box (TIR1/AFB; Mutte et al., 2018; Chapter 4). Biosynthesis pathway proteins include TAA family of amino transferase (TAA) and YUCCA family of monooxygenases (YUC; Chapter 3). It was also applied to the individual domains, Phox and Bem1 (PB1; Mutte and Weijers, 2020; Chapter 5), along with various downstream targets of the auxin pathway, such as SOSEKI (SOK; van Dop et al., 2020), Target of MOonopteros 5 (TMO5) and its interaction partner Lonesome HighWay (LHW; Chapter 3; Lu et al., 2020). Taken together, by following this protocol in combination with ever-growing high-quality sequence data, and leaping developments in the methods and algorithms in phylogenetics, reveal new evolutionary insights into our understanding of proteins and the crucial pathways.

Acknowledgements

The authors would like to thank the 1000 plant transcriptomes (OneKP) and Marine Micro Eukaryotic Transcriptome Sequencing Project (MMETSP) consortiums for providing such massive data resources for the scientific community. Efforts of all the authors is highly appreciated, who developed many extremely useful and efficient programs and algorithms for phylogenetics, and making them freely accessible to the scientific community.

Supplementary information

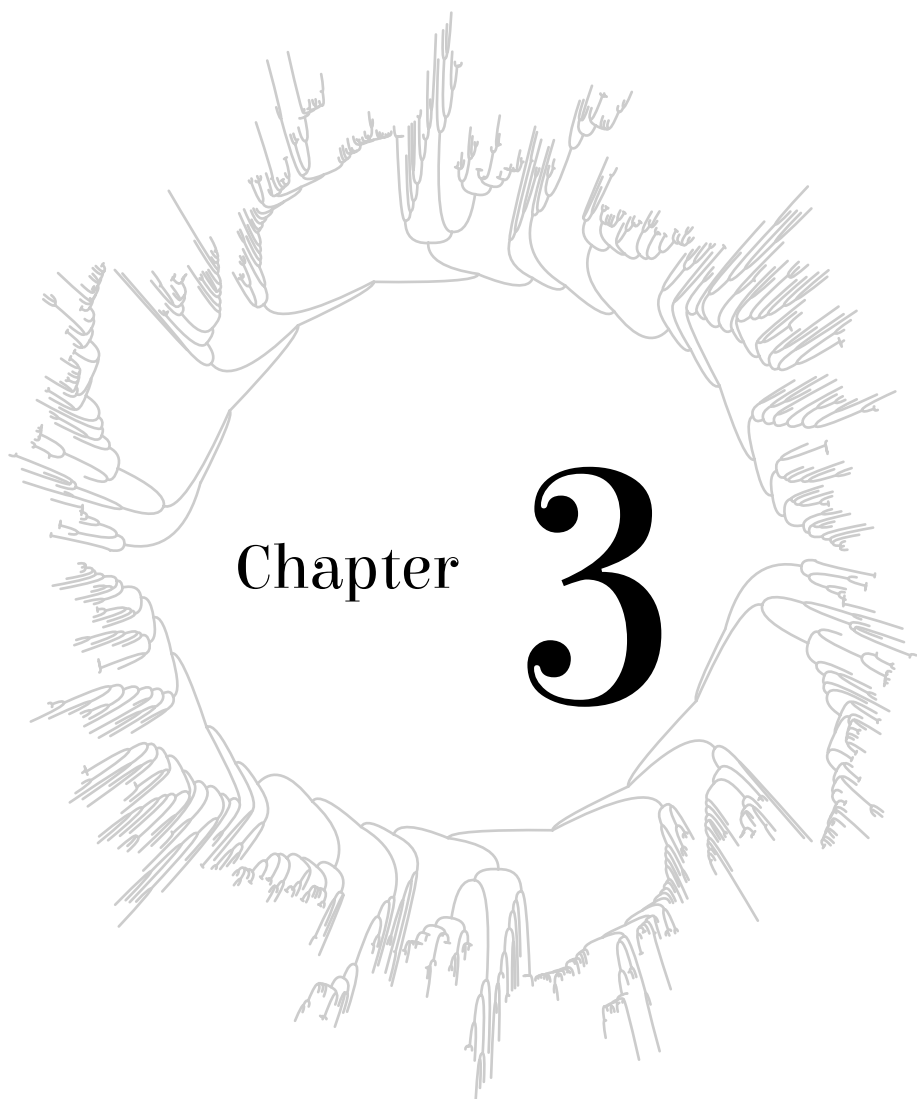
Recommended species to obtain query sequences from well annotated genomes

Plants	<i>Arabidopsis thaliana</i>	https://phytozome.jgi.doe.gov/pz/portal.html
	<i>Oryza sativa</i>	https://phytozome.jgi.doe.gov/pz/portal.html
	<i>Amborella trichopoda</i>	https://phytozome.jgi.doe.gov/pz/portal.html
	<i>Picea abies</i>	http://congenie.org/
	<i>Physcomitrella patens</i>	https://phytozome.jgi.doe.gov/pz/portal.html
	<i>Marchantia polymorpha</i>	https://phytozome.jgi.doe.gov/pz/portal.html
	<i>Chara braunii</i>	https://bioinformatics.psb.ugent.be/orcae/overview/Chbra
	<i>Klebsormidium nitens</i>	http://www.plantmorphogenesis.bio.titech.ac.jp/~algae_genome_project/klebsormidium/
Animals	<i>Homo sapiens</i>	UniProt: UP000005640
	<i>Mus musculus</i>	UniProt: UP000000589
	<i>Gallus gallus</i>	UniProt: UP000000539
	<i>Strongylocentrotus purpuratus</i>	UniProt: UP000007110
	<i>Caenorhabditis elegans</i>	UniProt: UP000001940
	<i>Drosophila melanogaster</i>	UniProt: UP000000803
Fungi	<i>Aspergillus nidulans</i>	https://mycocosm.jgi.doe.gov/mycocosm/home
	<i>Schizosaccharomyces pombe</i>	
	<i>Saccharomyces cerevisiae</i>	
	<i>Agaricus bisporus</i>	
	<i>Mortierella elongate</i>	
	<i>Rhizoclostridium globosum</i>	
Protozoa	<i>Dictyostelium discoideum</i>	UniProt: UP000002195
	<i>Entamoeba histolytica</i>	UniProt: UP000001926
	<i>Leishmania major</i>	UniProt: UP000000542
	<i>Monosiga brevicollis</i>	UniProt: UP000001357
	<i>Trypanosoma brucei</i>	UniProt: UP000008524

References

- Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* 37, 202–208.
- Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2019). The New Tree of Eukaryotes. *Trends Ecol. Evol.* In press, 1–13.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421.
- Carpenter, E.J., Matasci, N., Ayyampalayam, S., Wu, S., Sun, J., Yu, J., Jimenez Vieira, F.R., Bowler, C., Dorrell, R.G., Gitzendanner, M.A., et al. (2019). Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). *Gigascience* 8, 1–7.
- Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., and Flouri, T. (2019). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.* 1–4.
- van Dop, M., Fiedler, M., Mutte, S., de Keijzer, J., Olijslager, L., Albrecht, C., Liao, C.-Y., Janson, M.E., Bienz, M., and Weijers, D. (2020). A conserved biochemical paradigm underlies cell polarity across multicellular kingdoms. *Cell*. in press.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321.
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512.
- Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240.
- Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., and Jermini, L.S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589.
- Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
- Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMET-SP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* 12, e1001889.
- Lanfear, R., Calcott, B., Ho, S.Y.W., and Guindon, S. (2012). PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29, 1695–1701.
- Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259.
- Levy, S.E., and Myers, R.M. (2016). Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.* 17, 95–115.
- Li, F.W., Villarreal, J.C., Kelly, S., Rothfels, C.J., Melkonian, M., Frangedakis, E., Ruhsam, M., Sigel, E.M., Der, J.P., Pittermann, J., et al. (2014). Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proc. Natl. Acad. Sci. U. S. A.* 111, 6672–6677.
- Lu, K.-J., van 't Wout Hoffland, N., Mor, E., Mutte, S., Abrahams, P., Kato, H., Vandepoele, K., Weijers, D., and De Rybel, B. (2020). Evolution of vascular plants through redeployment of ancient developmental regulators. *Proc. Natl. Acad. Sci.* in press.
- Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., et al. (2014). Data access for the 1,000 Plants (1KP) project. *Gigascience* 3, 17.
- Mutte, S.K., Kato, H., Rothfels, C., Melkonian, M., Wong, G.K.-S., and Weijers, D. (2018). Origin and evolution of the nuclear auxin response system. *Elife* 7, e33399.
- Mutte, S.K., and Weijers, D. (2020) Deep Evolutionary History of the Phox and Bem1 (PB1) Domain Across Eukaryotes. *Sci Rep* 10, 3797.
- Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.
- One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685.
- Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542.
- Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. (2015). GUIDANCE2: Accurate detection of unreliable alignment

- regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* 43, W7–W14.
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313.
- Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656.
- Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577.
- Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J.C., and Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7.
- Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191.
- Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. U. S. A.* 111, E4859–E4868.
- Yerramsetty, P., Stata, M., Siford, R., Sage, T.L., Sage, R.F., Wong, G.K.-S., Albert, V.A., and Berry, J.O. (2016). Evolution of RLSB, a nuclear-encoded S1 domain RNA binding protein associated with post-transcriptional regulation of plastid-encoded *rbcl* mRNA in vascular plants. *BMC Evol. Biol.* 16, 141.
- Yoshida, S., van der Schuren, A., van Dop, M., van Galen, L., Saiga, S., Adibi, M., Möller, B., ten Hove, C.A., Marhavy, P., Smith, R., et al. (2019). A SOSEKI-based coordinate system interprets global polarity cues in *Arabidopsis*. *Nat. Plants* 5, 160–166.



Chapter 3

Reconstructing the evolutionary past of auxin biology

Sumanth Kumar Mutte, Rubaiat Nazneen Akhand and Dolf Weijers
Laboratory of Biochemistry, Wageningen University, Wageningen, the Netherlands

Parts of this chapter have been published as:

Lu, K.-J.^{*}, van 't Wout Hofland, N.^{*}, Mor, E., Mutte, S., Abrahams, P., Kato, H., Vandepoele, K., Weijers, D., De Rybel, B. (2020). Evolution of vascular plants through redeployment of ancient developmental regulators. *Proceedings of the National Academy of Sciences*, 117(1), 733–740. DOI: 10.1073/pnas.1912470117.

van Dop, M.^{*}, Fiedler, M.^{*}, Mutte, S., de Keijzer, J., Olijslager, L., Albrecht, C., Liao, C-Y., Janson, M., Bienz, M., Weijers, D. (2020). DIX Domain Polymerization Drives Assembly of Plant Cell Polarity Complexes. *Cell*, 180(3), 427–439. DOI: 10.1016/j.cell.2020.01.011.

^{*} These authors contributed equally



Auxin is a phytohormone involved in multiple processes controlling plant growth and development. The natural auxin Indole-3-acetic acid (IAA) is synthesized primarily through a Tryptophan-dependent pathway. In flowering plants, the expansion of biosynthesis gene families has allowed for dynamic regulation and tissue-specific expression of synthesis capacity. Beyond synthesis of auxin, conjugation of auxin with amino acids is an alternate approach to control levels of active hormone, thus contributing to auxin homeostasis. Despite the detailed understanding of enzymes and pathways regulating auxin biosynthesis and homeostasis, deep evolutionary insights into these gene families is lacking. In this study, we reconstruct the evolutionary history of the gene families involved in auxin biosynthesis (TAA and YUC) and homeostasis (GH3). Along with these, we also studied the evolutionary trajectory of the elusive auxin binding protein (ABP1), as well as of the transcriptional targets of the Nuclear Auxin Pathway i.e. TMO5, LHW and SOK. We found that gene families involved in both the auxin biosynthesis and homeostasis have evolved in the last common ancestor of all land plants. The GH3 family of amide synthases, that have multiple hormones as substrates, also showed early divergence of jasmonate- and auxin-specific clades in the common ancestor of land plants. We find that ABP1 has evolved before biosynthesis and genomic response components. Given that ABP1 likely mediates a non-transcriptional auxin response, these findings indicate that non-genomic auxin responses define an ancient auxin response system dating back to unicellular algae, while biosynthesis and genomic responses are specific to land plants.

Introduction

Auxin is an essential hormone in plant growth and development. The natural auxin, Indole-3-acetic acid (IAA), is synthesized through both tryptophan (Trp)-dependent and Trp-independent pathways (Brumos et al., 2014; Zhao, 2012, 2014). Indole synthase is a key enzyme in Trp-independent auxin biosynthesis (Wang et al., 2015). However, the exact role of indole synthase in auxin biosynthesis and its importance in plant growth is relatively unclear (Nonhebel, 2015). Of the various Trp-dependent routes of auxin biosynthesis, one pathway has been extensively studied. In this pathway, Trp is converted to IAA in two steps (Zhao, 2014). In the first step, TRYPTOPHAN AMINOTRANSFERASE OF ARABIDOPSIS (TAA) amino-transferases convert Trp to Indole-3-pyruvate (IPA). IPA is next converted to IAA by the YUCCA (YUC) family of flavin-containing monooxygenases (FMO). This two-step synthesis of IAA from Trp, the primary route of auxin biosynthesis in plants, has been explored in detail (Mashiguchi et al., 2011; Stepanova et al., 2008, 2011).

Once auxin is produced in the cell, its levels are tightly regulated to perform local functions (Zhao, 2018). Along with the controlled and localized production of auxin, inactivation through oxidation or conjugation is another important mechanism to maintain homeostasis (Ljung, 2013). Oxidation of IAA converts IAA to 2-oxindole-3-acetic acid (oxIAA). Furthermore, conjugation of IAA to amino acids is considered a storage mechanism, where IAA can be released from conjugates by hydrolysis (Ludwig-Müller, 2011; Zhang et al., 2017). IAA is converted to amide conjugates by GRETCHEN HAGEN 3 (GH3) amide synthases (Staswick et al., 2005). The *Arabidopsis* genome encodes 19 proteins of the GH3 family that differ in their substrate specificities. For example, Indole-3-acetyl-aspartic acid (IAA_{asp}) is formed through IAA conjugation with aspartate by the GH3.6 enzyme, and Indole-3-acetyl-glutamic acid (IAA_{glu}) is formed by conjugation to glutamate by GH3.17 (Staswick et al., 2005). Both IAA_{asp} and IAA_{glu} represent irreversible conjugates, leading to the inactive form of IAA (Östin et al., 1998). Interestingly, it was shown that some GH3 family members also have preference for other hormones, such as jasmonic acid (JA) or salicylic acid (SA) and the non-hormone benzoate. JASMONATE RESISTANT 1 (JAR1/GH3.11) is involved in the conjugation of JA to Isoleucine, forming the active JA-Ile form. The PPHB SUSCEPTIBLE 3 (PBS3/GH3.12) prefers benzoates and potentially SA as substrates over other hormones (Nobuta et al., 2007; Okrent et al., 2009; Staswick et al., 2002). Hence, the diversity in the substrate specificity of various GH3 family members evolved to maintain homeostasis of multiple hormones in plant cells.

Once auxin is synthesized and present in its active form, it elicits both genomic and non-genomic responses. Genomic or transcriptional responses by regulation of downstream gene transcription through the Nuclear Auxin Pathway (NAP) are relatively well understood (Weijers and Wagner, 2016). However, auxin also affects various cellular functions through non-genomic responses. As a part of this mechanism, one of the proteins that is relatively well explored is AUXIN BINDING PROTEIN 1 (ABP1; Woo, 2002). ABP1 localizes to the

endoplasmic reticulum (ER) and also to the apoplast. Through various immunological and transgenic approaches, as well as using a weak mutant allele, ABP1 was proposed to act in plasma membrane depolarization and also identified as an extracellular receptor, interacting with RHO OF PLANTS (ROP) and TRANSMEMBRANE KINASE 1 (TMK1), to alter the cell shape (Robert et al. 2010; Chen et al. 2014; Xu et al. 2014). Interestingly, when the ABP1 gene was mutated through CRISPR/Cas9 gene editing, none of these functions were affected, suggesting that ABP1 is not required for auxin signaling, plasma membrane depolarization or *Arabidopsis* development (Gao et al., 2015; Paponov et al., 2019). Therefore, ABP1 roles in auxin action, and in plant growth and development remain unclear.

Auxin-dependent genomic/transcriptional responses occur through a well-established NAP pathway, regulating numerous growth and developmental processes in plants (reviewed in Weijers and Wagner 2016). Some of the key downstream processes of NAP, among others, include the control of cell division orientation (Yoshida et al., 2014). This function may involve the SOSEKI (SOK) proteins that are activated by the AUXIN RESPONSE FACTOR 5 (ARF5)/MONOPTEROS protein and mark cell polarity (Yoshida et al., 2019). Furthermore, auxin response promotes vascular tissue development through the TARGET OF MONOPTEROS 5 (TMO5) and LONESOME HIGHWAY (LHW) basic Helix-Loop-Helix (bHLH) transcription factors (De Rybel et al., 2013). As auxin is widely present across distant phylogenetic lineages including bacteria (Cooke et al., 2002), it is worth investigating if the auxin pathway components and the downstream targets are also deeply conserved across plants and algae.

The evolution of key components of the NAP pathway is studied in detail (Chapter 4; Mutte et al., 2018). However, if the downstream targets of the pathway (SOK, TMO5 and LHW) also evolved along with the components of the pathway itself is yet to be elucidated. To reconstruct auxin biology at each transition in plant evolution, we performed a deep evolutionary analysis of components in auxin biosynthesis and homeostasis, ABP1, SOK, TMO5, and LHW families across land plants and algae. This has revealed that not only the biosynthesis components (TAA and YUC), but also the components of homeostasis (GH3) and downstream regulation (SOK, TMO5 and LHW) evolved in the early diverged land plants, but not in green algal ancestors (Zygnematophyceae). As the functional domains in SOKs are not yet annotated, we have also performed domain annotation based on the SOK homologs from land plants. This revealed novel protein domains and interesting evolutionary patterns. Thus, this study provides the evolutionary patterns of both the auxin biosynthesis and homeostasis gene families, and also the downstream components of auxin pathway, along with their functional domain annotations.

Results

Auxin biosynthesis pathway: TAA and YUCCA

The TAA(-related) gene family is represented by five homologs in *Arabidopsis* (TAA1, TAR1-4), which encode Alliinase domain-containing proteins. TAA1, TAR1 and TAR2 proteins (TAA clade) contain only the Alliinase-C domain (InterProID: IPR006948), whereas TAR3 and

TAR4 (Alliinase clade) contain both Alliinase-EGF (Epidermal Growth Factor; InterProID: IPR006947) and Alliinase-C domains, in the N- and C-terminus of the proteins, respectively. Enzymes in the TAA clade are considered core auxin biosynthesis enzymes, whereas TAR3 and TAR4 are considered Alliinases. First, we investigated the origin of both these homologous groups in the green algal lineage. Both chlorophytes and charophytes were identified with one ortholog in each species (Fig. 1).

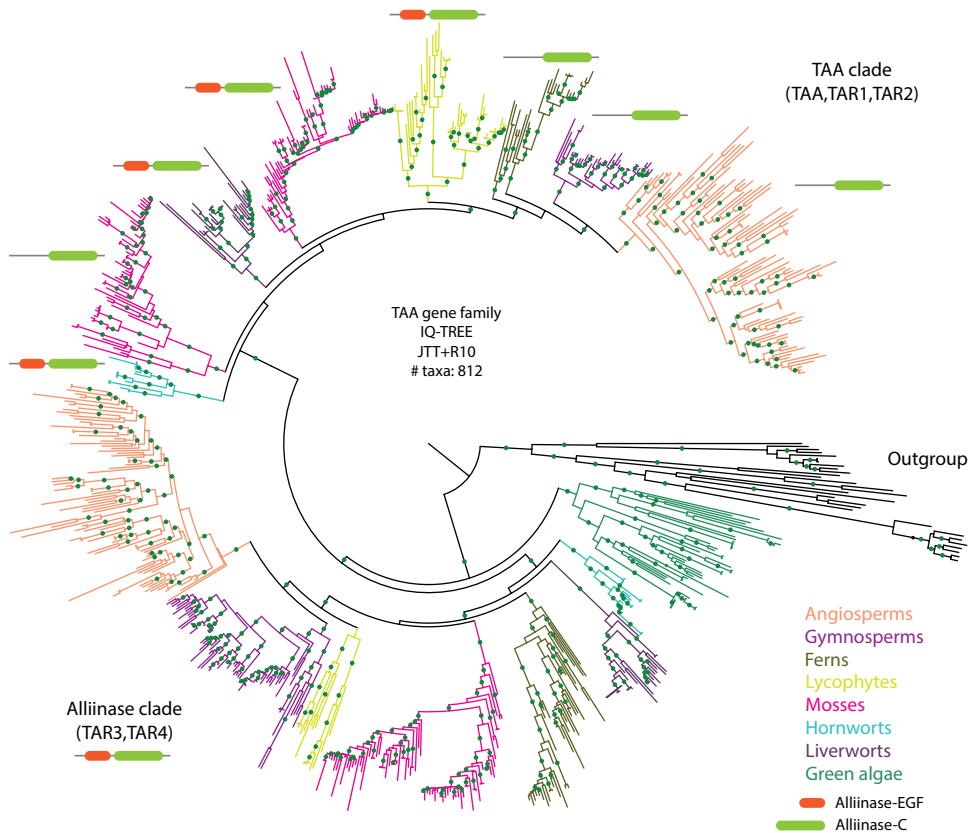


Figure 1: Phylogenetic tree of the TAA gene family with green algae and land plant homologs. Protein domains, Alliinase-EGF and Alliinase-C are indicated with ‘red’ and ‘green’ representations, respectively. TAR3/TAR4 clade shows the consistent presence of both domains in all lineages, whereas in TAA/TAR1/TAR2 clade, some phyla lack Alliinase-EGF domain. Aminotransferases from land plants other than TAA members were used as outgroup sequences to root the tree. Branches that are well-supported (bootstrap >75) are indicated in green dots. Orthologs from each phylum are represented with a different colour as indicated in the legend on the right bottom. Basic information about the tree construction: ‘software’, ‘model of evolution’ and the ‘number of taxa’ used for phylogenetic tree construction are indicated in the center. The complete tree can be found at interactive Tree of Life (iTOL): <https://itol.embl.de/shared/dolfweijers>.

Most of the chlorophyte orthologs identified had only the Alliinase-C domain, whereas the majority of the charophytes contain both the Alliinase-EGF and Alliinase-C domains. This indicates that having both domains is likely the ancestral state of TAA/Alliinase homologs. Among charophytes, Klebsormidiophyceae and Zygnematophyceae (closest algal ancestors of the

land plants) have TAA/Alliinase homologs, while these are absent from Charophyceae (Fig. 2). Among land plants, both the TAA and Alliinase clades are represented by a single ancestral copy across all the phyla from bryophytes until angiosperms. Both Alliinase domains were maintained in all phyla in the Alliinase clade. However, there was a loss of the Alliinase-EGF domain in one specific duplication in mosses and then consistently lost in ferns, gymnosperms and angiosperms in TAA clade. Hence, the lack of Alliinase-EGF domain in TAA1, TAR1 and TAR2 was not an ancestral loss, but rather happened later during the evolution of vascular plants.

The *Arabidopsis* genome encodes 11 YUC proteins. YUC orthologs in land plants can be divided into two sub-groups (Fig. 2 and Fig. 3). While all the 11 YUC homologs of *Arabidopsis* were placed in the YUC group, many other land plant (liverworts, hornworts, lycophytes and ferns) and green algae (charophytes) homologs were placed as sister clade to the YUC group, referred as sYUC group (Fig. 2A). Interestingly, the sYUC clade is lost in mosses and further discontinued in spermatophytes (Fig. 2). In a reciprocal BLAST search with the *Arabidopsis*

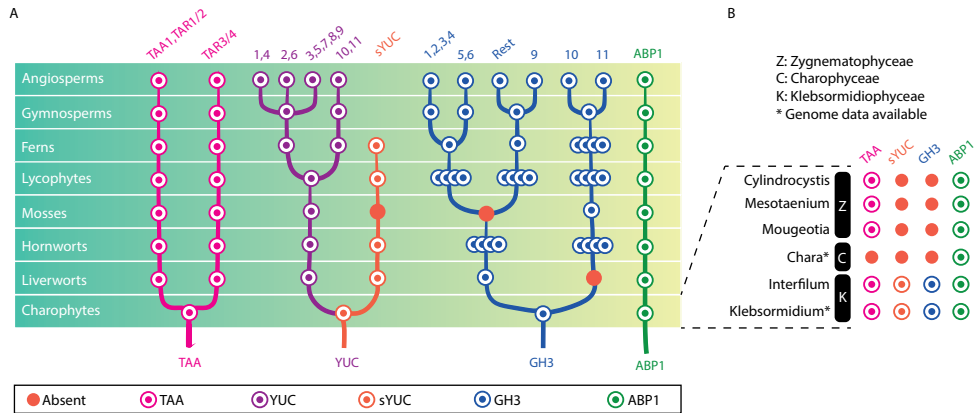


Figure 2: (A) Schematic summary of presence or absence of ancestral copy based on the well-supported clades in Figures 1, 2, 4 and S1. Each circle represents the presence of one ancestral copy in that phyla, for each gene family with a color represented below in the legend. Absence of an ancestral copy is represented with a filled 'red' circle. Lack of any circle or a filled dot in 'sYUC' clade represents the discontinuation of that clade further in spermatophytes. Multiple stacked circles represent the presence of multiple duplicate copies in that phylum. **(B)** Detailed inference of the presence and absence of each family in specific species covering major classes in charophytes. *Klebsormidium nitens* or *K. subtile*; *Interfilum paradoxum*; *Chara braunii*; *Mougeotia* sp.; *Mesotaenium caldariorum* or *M. kramstae* and *Cylindrocystis brebissonii* or *C. cushleackae*.

proteome, all the members of YUC and sYUC clades had hits with the respective YUC orthologs in *Arabidopsis*, confirming their close resemblance. As YUC proteins belong to a larger family of flavin-dependent monooxygenases (FMO), there is a possibility that the green algal ancestors identified here are common ancestors to other FMO family proteins as well, and not specific to the YUC family alone. To identify the ancestry, we included other FMO family proteins such as FLAVIN MONOOXYGENASE 1 (FMO1) and FMO GLUCOSINOLATE S-OXYGENASE (GS-OX) in the phylogenetic analysis. This analysis clearly differentiated the YUC orthologs in land plants and charophytes from other related FMO orthologs (Fig. 3). A similar sister clade

like sYUC was observed as sister to FMO family members, referred as sFMO. Likewise, this gene was also lost in seed plants, and independently in lycophytes (Fig. 2A and Fig. 3). The reciprocal BLAST of all FMO family members with the *Arabidopsis* proteome confirmed the orthologous relation with FMO family members, rather than YUC family proteins. Surprisingly, all the members in the sFMO clade showed best hits with YUC family members, not with FMO family proteins. However, the detailed origin of these YUC and FMO clades is unclear as very few homologs were identified in green or red algae in the sFMO clade and also no red algae in the (s)YUC clade (Fig. 2B and Fig. 3B). Among land plants, YUC orthologs were maintained as a single ancestral state up to lycophytes, while a first duplication in ferns gave rise to the YUC10/11 group that was further retained in gymnosperms and angiosperms. On the other hand, YUC1/4, YUC2/6 and YUC3/5/7/8/9 formed three sub-groups in angiosperms (Fig. 2A). No phylum specific deletions were observed in any of the sub-groups of YUCs in land plants. It is worth noting that the green algal ancestors of land plants i.e. Zygnematophyceae, and also Charophyceae lack YUC orthologs (Fig. 2B).

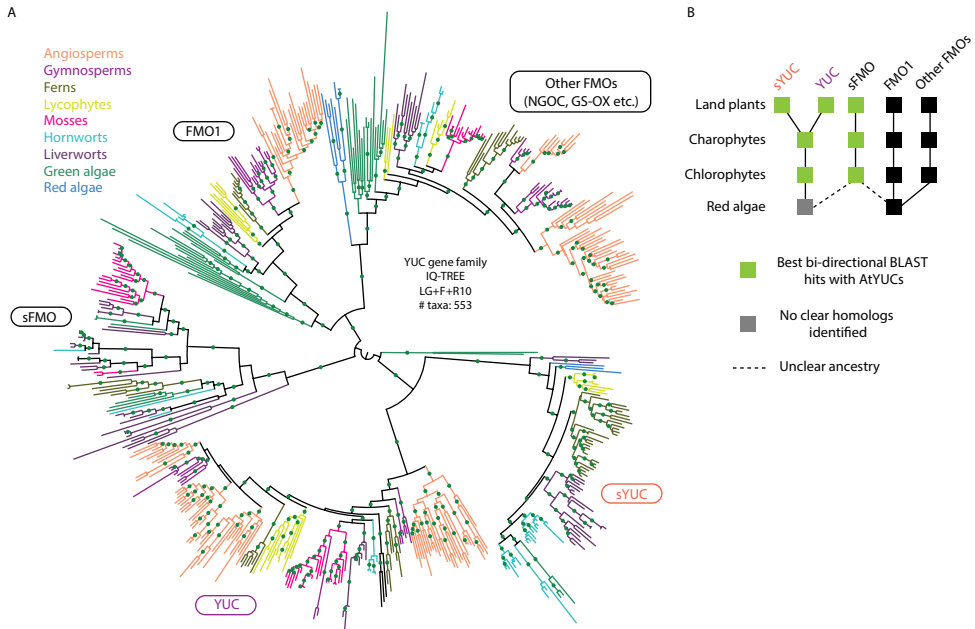


Figure 3: (A) Phylogenetic tree of the YUC gene family with algae and land plant homologs, along with other FMO families. Other information is similar to Figure 1. (B) Simple schematic showing the ancestry of algal YUC and FMO homologs. 'Green' squares represent the presence of homologs with best bidirectional BLAST with *Arabidopsis* YUC family members. 'Grey' squares represent the lack of clear homologs. 'Black' squares represent the presence of FMO homologs with respective hits to the FMO families. 'Continuous lines' represent the clearly inferred ancestry and 'dotted lines' represent the unclear ancestry.

Homeostasis: GH3

The *Arabidopsis* genome contains 19 members of the GH3 gene family. All these 19 members could be divided into 6 clades in the angiosperms (Fig. 2A and Fig. 4). All these 6 clades emerged

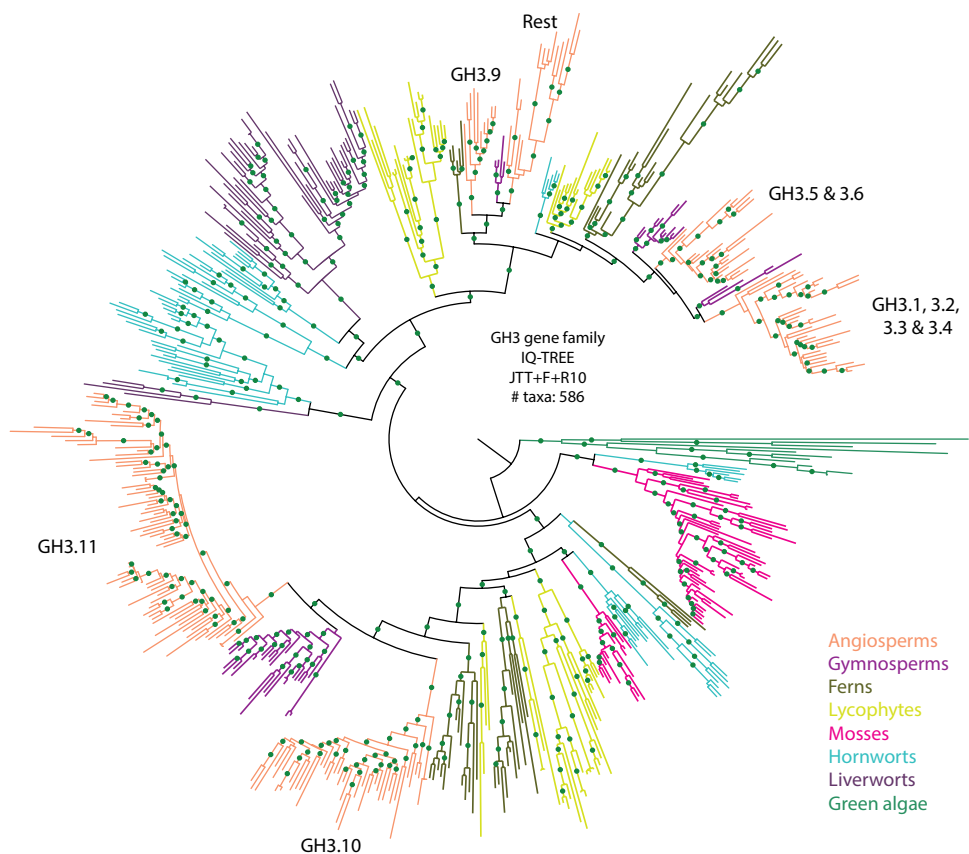


Figure 4: Phylogenetic tree of the GH3 gene family with green algae and land plant homologs. Respective *Arabidopsis* orthologs that are present in the specific clade are mentioned with the corresponding *Arabidopsis* family members. ‘Rest’ includes GH3.7, 3.8, 3.12 until 3.19. See Figure 1 legend for other information.

from a single ancestral charophyte copy, which was observed only in Klebsormidiophyceae, but not in Charophyceae and Zygnematophyceae, similar to the YUC family (Fig. 2B). The GH3 family was split into two clades in land plants, that were represented by multiple copies in hornworts in both clades. Among the bryophytes, the GH3.10/11 clade is lost in liverworts, whereas the other clade was lost in mosses (Fig. 2A and Fig. 4). Interestingly, similar to hornworts, multiple duplicates were identified in both clades in lycophytes. As the phylogenetic analysis was based on transcriptomes, there is a possibility that the duplicates arose due to mis-assembly of the multiple transcripts of the same gene. To confirm this, we have checked for orthologs in *A. agrestis* and *S. mollendorffii* genomes, and indeed found that these are legitimate duplicates, with more than 20 copies of GH3-encoding genes in each genome. Interestingly, no such duplications have been identified in liverworts and mosses, with only two copies being encoded in each genome of both *M. polymorpha* and *P. patens*. In ferns and gymnosperms, the GH3.10/11

clade is kept intact, whereas further duplications were observed in the other clade, giving rise to four sub-groups in angiosperms: GH3.1/2/3/4 – GH3.5/6 – GH3.9 and the rest of the GH3 proteins i.e. GH3.7/8, GH3.12 till GH3.19 (Fig. 2A and Fig. 4). Due to the abnormal nature of duplications in hornworts and lycophytes, it was difficult to interpret the ancestral copy in these highly duplicated groups. Hence, genome based synteny information is required to differentiate the ancestral states from phylum specific duplications.

Non-genomic responses: ABP1

ABP1 is a single copy gene identified in all the phyla of land plants as well as green algae (Fig. 2A and Fig. S1A). This well conserved gene family showed no ancestral duplications or losses in any of the phyla. Interestingly, it was lost only in the class Marchantiopsida of the liverworts, including the model liverwort *M. polymorpha*. Moreover, this is the only gene family that was present in all the three major classes of charophytes i.e. Klebsormidiophyceae, Charophyceae and Zygnematophyceae (Fig. 2B). Based on the structure of maize ABP1 in complex with 1-Naphtheleneacetic acid (1-NAA; synthetic auxin; Woo, 2002), we checked if the residues needed for auxin binding are also deeply conserved (Fig. S1B). Indeed, the amino acids in the auxin binding pocket are deeply conserved in chlorophytes, whereas the Zinc binding residues are deeply conserved even in red algae (Fig. S1B).

Targets of NAP: TMO5, LHW and SOSEKI

TMO5 and LHW are basic Helix-Loop-Helix (bHLH) transcription factors. The TMO5 gene family is represented by five members in *Arabidopsis*, whereas LHW is represented by four members (Fig. 5). Both these gene families have unique ancestors in charophytes, especially in the Zygnematophyceae, the algal sisters of land plants. No phylum-specific losses could be identified either in the TMO5 or LHW gene families (Fig. 5). LHW homologs were not found in the genome of *P. patens*, but identified in other mosses, indicating an independent loss in *P. patens*. The first duplication in the TMO5 gene family was identified in gymnosperms, whereas in LHW, the first ancestral duplication occurred only in angiosperms, which divided the family into three sub-groups (Fig. 5). Since bHLH proteins represent much larger gene families in land plants, we used the closest bHLH orthologs of TMO5 and LHW from several land plants as an outgroup in the phylogeny. The results confirmed that all TMO5 and LHW orthologs identified were indeed true TMO5/LHW orthologs (refer iTOL tree).

SOSEKI family is represented by five members in *Arabidopsis*. In this study, we found the SOK proteins to be limited to land plants, and absent from all algal sister groups (Fig. 5). A single SOK ancestor was found until a first duplication in the ancestor of ferns. Subsequent duplications in seed and flowering plants increased the number of homologs (Fig. 5 and 6A). As there are no annotated domains known in SOK proteins, we mined the full set of land plant SOSEKI sequences for conserved motifs. We identified common domain topology, an N-terminal domain with superficial resemblance to animal DIX domains (Yoshida et al., 2019).

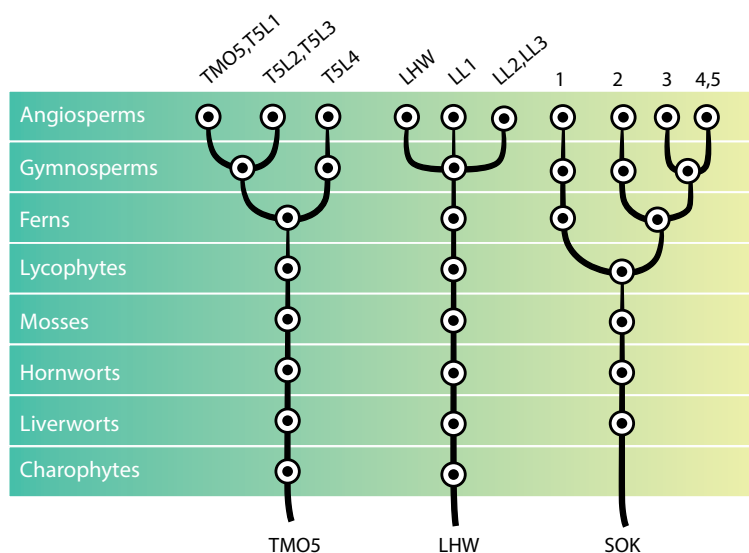


Figure 5: Schematic summary of ancestral copies. Based on the well-supported clades in the respective phylogenetic trees for TMO5, LHW and SOK proteins. Each circle represents the presence of one ancestral copy in that phyla, for the respective gene family. Full version of the complete phylogenetic trees for all the three families can be found in iTOL: <https://itol.embl.de/shared/dolfweijers>. T5L1-4, TMO5-LIKE1-4; LL1-3, LHW-LIKE1-3.

Further downstream, SOK proteins contain a TQT motif (Fig. 6B) and a CG motif (Fig. 6C). Within their C-termini, they exhibit a C2HC zinc finger (ZnF) signature, found in bryophyte and lycophyte SOSEKI and in the SOK1 clade, while the other clade, consisting of SOK2-5 orthologs of vascular plants, is characterized by a central KEY motif (Fig. 6D and 6E).

Although the DIX domain was originally thought to be limited to the Wnt pathway (Dishevelled, Axin and Dixin proteins) in animals (Dillman et al., 2013), our analysis suggests that a homolog of this domain is present throughout land plants (Fig. 6A). We thus extended our phylogenetic analysis beyond animals and plants, to study DIX domain-containing proteins in other eukaryotic lineages. We used fungal genomes (MycoCosm) and unicellular eukaryote transcriptome datasets (MMETSP) to search for DIX-like sequences. While our searches of the fungal kingdom were negative, we identified DIX-like sequences in organisms of the SAR (Stramenopiles, Alveolates and Rhizaria) group (Fig. 6A). Interestingly, none of the DIX-like sequences in SAR group organisms are associated with any of the conserved domains found in plant SOSEKI or animal DIX domain-containing proteins (Fig. 6A), suggesting that they may have different functional contexts.

Discussion

Phylogenetic analysis provides detailed evolutionary insights into the origin and diversification of auxin biosynthesis components. The two key enzymes in Trp-dependent auxin biosynthesis pathway, TAA and YUC, are deeply conserved across all the land plants. Evolution of TAA proteins was under debate regarding the ancestral state, where the charophyte algae (*K. nitens*)

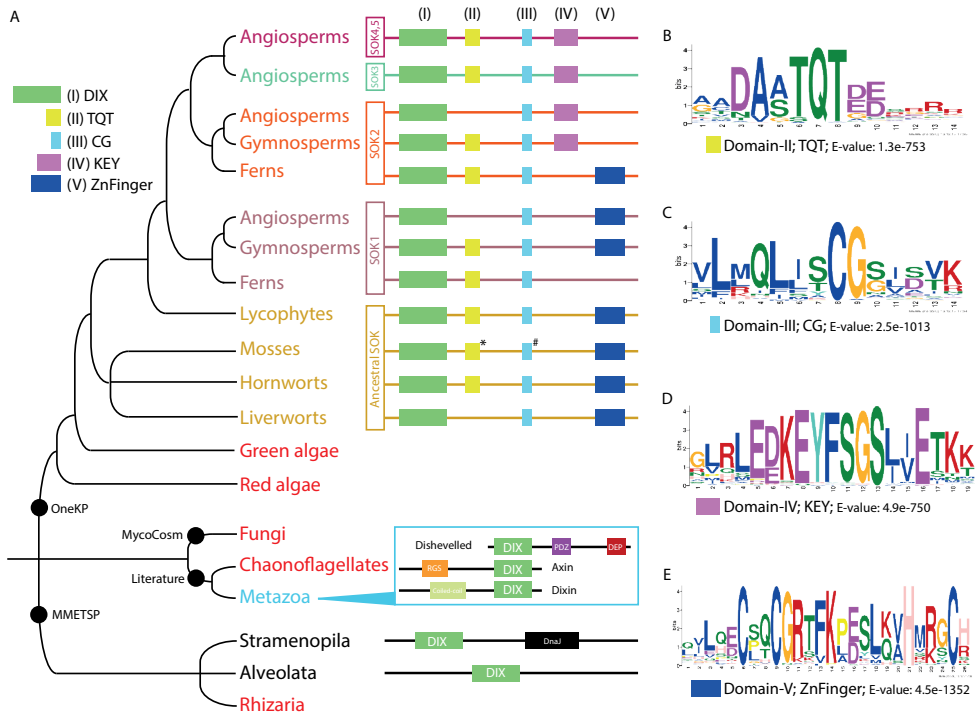


Figure 6: (A) Extended phylogeny and conserved domain topology of land plant SOSEKI proteins and DIX domain-containing proteins across eukaryotic kingdoms. * or # represent the degenerated motif. Sources of data is indicated with 'black' filled circles (OneKP; MycoCosm; MMETSP; Literature). boxes: domain topology. 'Red' branches i.e. Green algae, Red algae, Fungi, Chaenoflagellates and Rhizaria were identified to lack DIX domain-containing proteins. **(B-E)** Amino acid logos of conserved motifs in plant SOSEKI proteins.

were assumed to have ancestors for Alliinases (TAR3/TAR4) but not TAA orthologs (Turnaev et al., 2015; Wang et al., 2016). However, based on few species, it was later shown that there is a TAA-like gene in Zygnematophyceae, the algal ancestors of land plants (Romani, 2017), but still the domain architecture of these genes was unclear. Based on the current results, we have established that there is one single ancestor that gave rise to both TAA/Alliinase orthologs in land plants (Fig. 1 and Fig. 2A).

TAA orthologs (TAA1, TAR1 and TAR2) are known to be involved in auxin biosynthesis, thus contributing to plant growth and development (Stepanova et al., 2008). However, no biochemical studies have shown the role of Alliinases (TAR3 or TAR4) in auxin biosynthesis. Moreover, they are considered as Alliinases because they have an Epidermal Growth Factor (EGF) domain that majority of the Alliinase (and Alliinase-like) proteins have, but not TAA proteins. EGF has small di-sulfide rich β -strands that are important for binding to other proteins (Kuettner et al., 2002). Surprisingly, we found that bryophyte and lycophyte TAA orthologs have an EGF domain, similar to TAR3/TAR4 Alliinases, but the EGF domain is lost later in TAA orthologs. Functional analysis of the liverwort *M. polymorpha* TAA ortholog (MpTAA)

showed that, despite the presence of EGF domain, MpTAA functions in a similar way to its *Arabidopsis* counterpart, AtTAA1 (Eklund et al., 2015). Hence, the role of EGF domain in auxin biosynthesis is unclear. Given that EGF domains are involved in mediating protein interactions (Campbell and Bork, 1993; Kuettner et al., 2002), it is worth identifying the significance of the presence of EGF domain in bryophyte TAA orthologs in protein-protein interactions.

taa1/tar2 mutants fail to develop basal part of the embryo, where *yuc1/yuc4/yuc10/yuc11* quadruple mutants also showed a similar phenotype, indicating their importance in the same tissue with common expression patterns (Cheng et al., 2007; Stepanova et al., 2008). YUC10/11 was split early from the rest of the YUC orthologs in the common ancestors of ferns and seed plants. Another study has shown that *yuc1/yuc4* double mutants make few floral organs, whereas, *yuc1/yuc4* double, *yuc1/yuc2/yuc4* triple and *yuc1/yuc2/yuc4/yuc6* quadruple mutants produce fewer veins, indicating their important role in vascular formation (Cheng et al., 2006). Interestingly, both the shoot-functional YUC1/4 and YUC2/6 originated from the same common ancestor of all angiosperms, along with the root-functional YUC3/5/7/8/9. The quintuple mutants of these root-functional homologs (*yucQ*) leads to development of short and agravitropic roots (Chen et al., 2014). Hence, the tissue-specific functional differentiation in the shoot and root specific expression patterns can be considered as a flowering plant innovation to cope with the tissue and cell-type specificities in these complex species.

In the non-seed plants (except mosses), a novel clade (sYUC) was identified that is derived from the common ancestor of all land plants, and sister to all YUC orthologs. This clade was kept intact without any ancestral duplications until its loss in spermatophytes. Hence, it is likely that sYUC orthologs evolved to provide an alternate YUC function, before further duplications in the core YUC sub-groups. In our previous study, we have identified similar components (ncARF) that diverged from the canonical core ARF orthologs, that nonetheless act as positive regulators in auxin-dependent transcription (Mutte et al., 2018). It will be interesting to see if sYUC proteins are catalytically active, and whether they likewise act as positive components in auxin biosynthesis.

Despite the tissue- or cell-type-specific expression patterns of IAA biosynthesis genes, regulated inactivation of IAA also plays an important role in dynamic regulation and homeostasis (Ludwig-Müller, 2011). GH3 proteins that are important for maintaining hormone homeostasis are not only specific for auxin but some can also conjugate JA, SA and other benzoates (Westfall et al., 2010). Among the 19 GH3 homologs encoded in *Arabidopsis* genome, hormonal specificities for many are still unknown. GH3.2-6 and GH3.17 favor auxin, whereas, GH3.11 favors JA (Staswick et al., 2005). This coincides with the evolutionary patterns of divergence, where GH3.10/11 were split early from the rest in the common ancestor of land plants, where this clade could have a preference specifically for JA, while the other clade might have a broader preference for other hormones. Looking into the specific amino acids in either nucleotide or hormone binding sites revealed that certain amino acid positions in hormone binding site have specific preferences (Fig. S2). For example, GH3.10-11 have preference for a polar, acidic and

tryptophan residues in 130, 337 and 344 positions, respectively. Whereas, these positions prefer positive and hydrophobic residues in other clade, indicating different preference of hormones (Fig. S2). In contrast to the inactivating role that GH3 enzymes play in auxin activity, the formation of JA conjugates by GH3 enzymes is important for signaling. The JA receptor COI1 recognizes a conjugate of JA with Isoleucine (JA-Ile; Katsir et al. 2008). We found that the GH3 family in hornworts, lycophytes and ferns underwent numerous duplications that were not observed in other auxin-related gene families. It is unclear if this is a mechanism adopted by these phyla to maintain the hormonal homeostasis or if this is a part of block duplicates along with many other gene families. Further investigation into synteny based on the genomes may resolve this discrepancy.

Even though the dedicated auxin receptor TIR1 appeared only in land plants, it is interesting to find that ABP1 has even deeper roots in evolution, and is represented in chlorophytes and even red algae. Interestingly, the amino acids in the auxin binding pocket are deeply conserved even in chlorophytes (Fig. S1B). This indicates a possibility that the auxin-binding capacity of ABP1 evolved in chlorophytes, meaning that potential auxin-dependent non-genomic responses through ABP1 could have originated earlier than the genomic responses through TIR1, and auxin biosynthesis. It is worth noting that, despite the presence in all phyla, ABP1 may not be mandatory for survival and development, as it was lost in some species, including *Marchantia polymorpha*. Further genetic analysis of remote homologs of ABP1 may help resolve its role in auxin biology.

In summary, the trends observed in both biosynthesis and homeostasis indicate their origin and existence in land plants, but not in charophytes, especially Zygnematophyceae. In a similar way, one of the auxin output components, SOK, is also limited to land plants (van Dop et al., 2020). This is further supported by the finding that TMO5 and LHW orthologs from most land plants are functionally relevant for vascular development, while the *Klebsormidium* ortholog is not (Lu et al., 2020). Taken together, it is evident that the auxin pathway could be functional only in land plants, but not in the algal ancestors of land plants.

Materials and methods

Phylogenetic tree construction

For the phylogenetic tree construction of each gene family studied in this chapter, a similar methodology was used as described earlier (Chapter 2). Only the differences in key parameters that deviate from the default settings are mentioned here. After collecting the homologous sequences for each gene family, MAFFT iterative refinement algorithm (E-INS-i) was used to align the protein sequences. Alignment positions with more than a certain percentage of gaps specific to each gene family, as mentioned in the table below, were removed before the phylogeny construction. The most suitable evolutionary model for all the gene families as predicted by the ModelFinder, along with the number of taxa used for phylogenetic inference are given in the table below. The maximum-likelihood algorithm implemented in IQ-TREE was used for tree

construction. Obtained trees were visualized using the iTOL phylogeny visualization program. Phylogenetic trees were cleaned up manually for misplaced sequences as well as for clades with long branch attraction.

Gene family	% gaps removed	Evolutionary model	# taxa
TAA	20	JTT+R10	812
YUC	30	LG+F+R10	553
GH3	20	JTT+F+R10	586
ABP1	20	WAG+R8	347
TMO5	20	JTT+R7	126
LHW	40	JTT+F+R5	111
SOK	40	JTT+F+R8	207

For the GH3 family, since a very high redundancy was observed in non-seed plants, CD-HIT (Fu et al., 2012) was used to remove redundant sequences that are identical up to 90% and above, as they do not carry any novel phylogenetic signal. To extract the non-plant SOSEKI proteins or DIX domains, the BLAST module at JGI MycoCosm (genome.jgi.doe.gov/fungi) was used to search for DIX domain-containing proteins in fungi with plant (*A. thaliana*) and animal (*H. sapiens*) DIX domains as query sequences. To determine the presence and evolution of SOSEKI or DIX domain-containing proteins in the SAR group, the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) data (Keeling et al., 2014) was used, and homologous genes were identified as described earlier (Chapter 2). The complete trees can be found at interactive Tree of Life (iTOL): <https://itol.embl.de/shared/dolfweijers>.

Identification of domains and motifs in SOSEKI proteins

Protein sequences used in phylogenetic tree construction were used for domain finding using the MEME motif discovery program (v4.12.0) with additional parameters “-mod zoops -nmotifs 15 -minw 10” (Bailey et al., 2009). Among 15 elements identified, 4 spanned the N-terminal 100 residues, and were identified together as DIX domain. Motifs that were specific to a certain clade or motifs that did not show conservation of significant amino acids were discarded.

Acknowledgment

We thank Dr. Joao Ramalho for helpful comments on the manuscript. This research was supported by a VICI grant to DW, from the Netherlands Organization for Scientific Research (NWO; 865.14.001).

Supplementary information

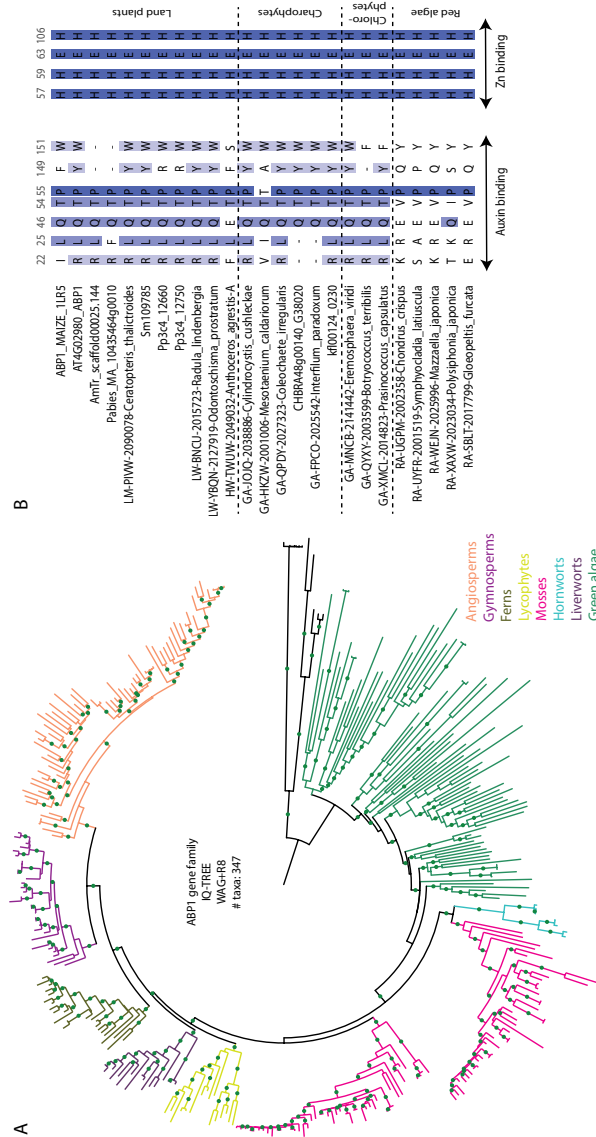


Figure S1: (A) Phylogenetic tree of the ABP1 gene family with green algae and land plant homologs. Branches that are well-supported (bootstrap >75) are indicated in green dots. Orthologs from each phylum are represented with a different colour as indicated in the legend on the right bottom. Basic information about the tree construction: 'software', 'model of evolution' and the 'number of taxa' used for phylogenetic tree construction are indicated in the center. **(B)** Deep conservation of key amino acids in auxin binding pocket as well as Zinc binding site of ABP1 proteins. Auxin (1-Naphthaleneacetic acid; 1-NAA) binding residues are conserved in chlorophytes and Zinc binding residues are conserved even in red algae. Light blue to dark blue color gradient represents low to high conservation, respectively. Numbering on the top is based on maize ABP1 protein (PDB ID: 1LR5; Woo, 2002).

	109	119	314	338	342	360	427	443	446	121	130	134	137	174	175	179	231	312	337	344	
AT4G27260_GH3.5_WES1	SSGTS	GGER	KK	T	YASS	E	D	FIC	R	L	R	Y	L	VL	Y	V	I	M	Y		
AT5G54510_GH3.6_DFL1	SSGTS	GGER	KK	T	YASS	E	D	FIC	R	L	R	Y	L	VL	Y	V	I	M	Y		
OS01G57610	SSGTS	GGER	KK	T	YASS	E	D	FVRR		M	R	Y	L	VL	Y	V	I	M	Y		
ATR_00043G00720	SSGTS	GGER	KK	T	YASS	E	D	FVCR		M	R	Y	L	VL	Y	V	I	M	Y		
AT1G59500_GH3.4	SSGTS	AGER	KK	T	YASS	E	D	F1RR		I	R	G	L	AL	Y	V	I	I	Y		
AT2G23170_GH3.3	SSGTS	AGER	KK	T	YASS	E	D	FVRR		M	R	Y	L	VL	Y	V	I	M	Y		
AT4G37390_GH3.2_BRU6	SSGTS	AGER	KK	T	YASS	E	D	F1RR		L	R	Y	L	VL	Y	V	I	M	Y		
AT2G14960_GH3.1	SSGTS	AGER	KK	T	YASS	E	D	FVRR		L	R	Y	L	VL	Y	V	I	M	Y		
OS07G40290	SSGTS	AGER	KK	T	YASS	E	D	FVRR		L	R	Y	L	VL	Y	V	I	M	Y		
OS01G55940	SSGTS	AGER	KK	T	YASS	E	D	FVRR		L	R	Y	L	VL	Y	V	I	M	Y		
ATR_00001G05460	SSGTS	AGER	KK	T	YASS	E	D	FVRR		L	R	Y	L	VL	Y	V	I	M	Y		
LM-WGTU-2007729-Leucostegia_immersa	SSGTS	SGGK	KK	S	YAGS	E	D	LVGR		M	R	G	D	AS	Y	L	A	F	C		
LM-BMJR-2063067-Adiantum_tenerum	SSGTS	AGER	KK	T	YASS	E	D	FVCR		F	R	Y	L	VL	Y	V	I	M	Y		
AT5G13320_GH3.12_PB53	SSGTS	GGAQ	KK	T	YASS	E	D	FVRR		L	L	Y	R	AT	Y	I	V	T			
AT5G13370_GH3.15	SSGTS	GGVP	KK	T	YASS	E	D	FVGR		L	R	S	Y	MV	F	P	I	F	F		
AT1G23160_GH3.7	SSGTS	GGAQ	KK	T	YASS	E	D	F1RR		L	L	Y	R	AS	Y	I	I	I	Q		
OS11G32520	SSGTS	GGQK	PK	T	YASS	E	D	-FPT		L	K	Y	L	VL	Y	V	I	M	Y		
ATR_00016G02310	SSGTS	GGHP	KK	T	YASS	E	D	FVRR		L	K	Y	L	VL	Y	V	I	M	Y		
AT2G47750_GH3.9	SSGTS	AGER	KK	T	YASS	E	D	FICR		L	R	Y	L	VL	Y	V	I	M	Y		
OS07G47490	SSGTS	RGE	PK	T	YASS	E	D	FICR		L	R	Y	L	VL	Y	V	I	M	Y		
ATR_00045G02570	SSGTS	GGES	KK	T	YASS	E	D	FICR		I	R	Y	L	VL	Y	V	I	M	Y		
Mapoly005350073	SSGTS	GGDP	KK	T	YAS	T	E	FMRR		K	R	W	T	AV	F	I	I	L	Y		
Mapoly004250030	SSGTS	GGCK	KK	T	YASS	E	D	FLGR		R	L	F	I	VS	F	L	I	V	A		
AT2G46370_GH3.11_JAR1	SSGTS	GGRP	KK	T	YASS	E	D	FICR		M	T	F	A	AT	V	V	I	D	W		
AT4G03400_GH3.10_DFL2	SSGTS	TGER	KK	T	YGS	T	E	F1YR		A	T	F	S	AT	V	A	I	D	W		
OS05G50890	SSGTS	THGR	KK	T	YGAS	E	D	FICR		L	T	Y	S	AT	L	T	I	D	W		
OS01G12160	SSGTS	TGKR	KK	T	YGAS	E	D	FVCR		V	T	Y	S	AT	V	T	I	E	W		
ATR_00093G00010	SSGTS	TGKR	KK	T	FGSS	E	D	FVCR		L	T	F	A	AT	V	T	I	D	W		
ATR_00093G03200	NSGTS	TGER	KK	T	YGAS	E	D	FVGR		A	S	F	A	AT	Y	T	I	E	W		
Pp3c24_16260	SSGTS	TAGK	PK	S	YAA	C	E	D	FVRR		Y	F	Q	V	QS	H	T	V	A	F	
Pp3c10_20960	SSGTS	TGCK	CK	S	YGAS	E	D	FVCR		L	A	G	G	GT	F	I	I	D	W		
GA-FPCO-2003223-Interfilum_paradoxum	TSGS	SSGTQ	KK	T	YAA	S	E	D	YFYR		A	L	G	F	ST	G	N	V	H	P	
GA-FPCO-2003225-Interfilum_paradoxum	SSGS	SSGTQ	KK	T	YVA	S	E	D	YFYR		S	A	S	F	AT	G	N	V	L	T	
GA-FPCO-2028952-Interfilum_paradoxum	SSGTS	TGDR	-K	T	YAC	S	E	D	FQHR		M	V	D	T	IV	L	T	I	T	L	
kfl00471_0050	TSGS	TDS	R-K	T	YGS	S	E	D	FQYR		L	A	I	A	LV	L	S	I	T	F	
kfl00006_0200	SSGTS	TGR	-K	T	YAC	S	E	D	FQYR		M	S	G	T	IV	L	T	I	T	L	

← Nucleotide binding site →

← Acyl acid binding site →

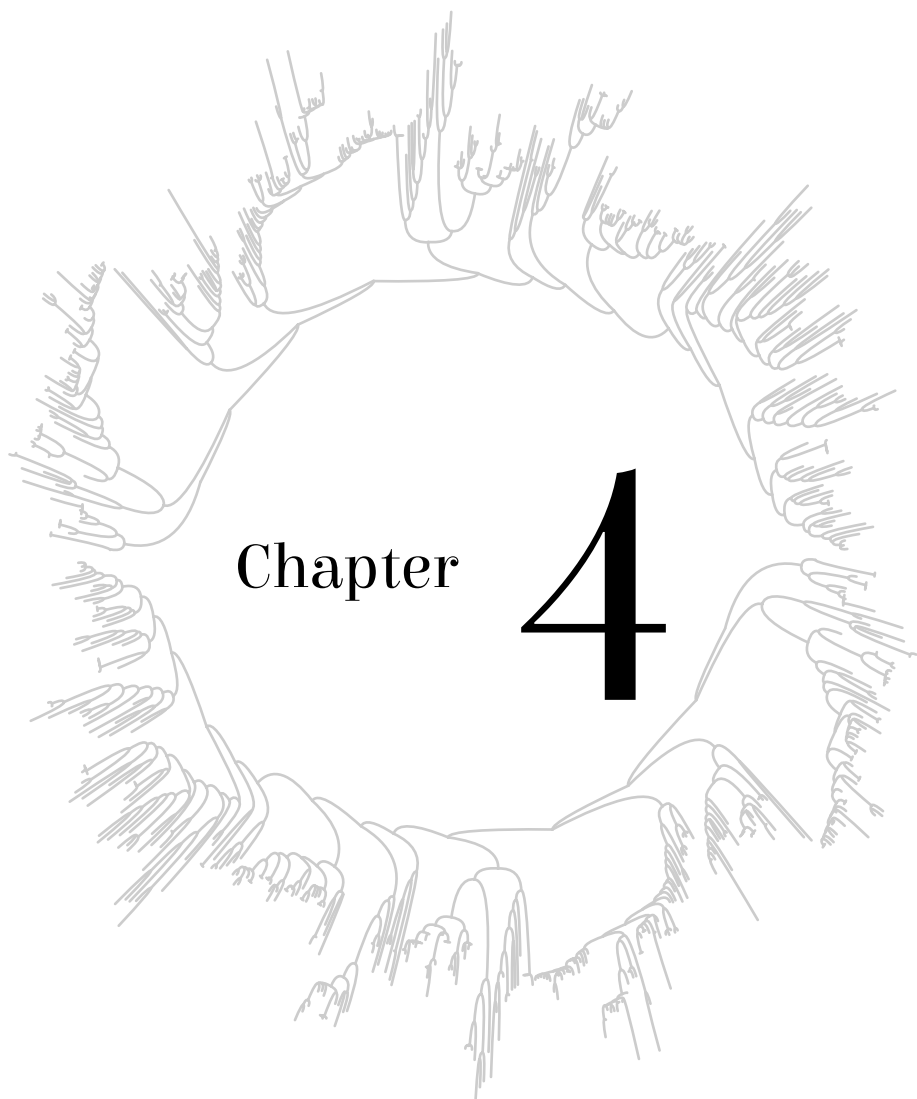
Figure S2: Conservation of key amino acids in nucleotide and acyl binding pockets of GH3 proteins. Orthologs from various species in each clade they belong to, is indicated in the right. Amino acids are colored according to their property: Blue, Positive; Red, Negative; Orange, Hydrophobic, Grey, Polar. Numbering on the top is based on Arabidopsis GH3.5 protein sequence. Important amino acids in various hormone binding pockets are based on Westfall et al., 2012, 2016. Starting letters of identifiers represent the species names: AT, *Arabidopsis*; OS, Rice; ATR, *Amborella*; Mapoly, *Marchantia*; Pp, *Physcomitrella*; kfl, *Klebsormidium*.

References

- Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* 37, 202–208.
- Brumos, J., Alonso, J.M., and Stepanova, A.N. (2014). Genetic aspects of auxin biosynthesis and its regulation. *Physiol. Plant.* 151, 3–12.
- Campbell, I.D., and Bork, P. (1993). Epidermal growth factor-like modules. *Curr. Opin. Struct. Biol.* 3, 385–392.
- Chen, Q., Dai, X., De-Paoli, H., Cheng, Y., Takebayashi, Y., Kasahara, H., Kamiya, Y., and Zhao, Y. (2014a). Auxin Overproduction in Shoots Cannot Rescue Auxin Deficiencies in Arabidopsis Roots. *Plant Cell Physiol.* 55, 1072–1079.
- Chen, X., Grandont, L., Li, H., Hauschild, R., Pague, S., Abuzeineh, A., Rakusová, H., Benkova, E., Perrot-Rechenmann, C., and Friml, J. (2014b). Inhibition of cell expansion by rapid ABP1-mediated auxin effect on microtubules. *Nature* 516, 90–93.
- Cheng, Y., Dai, X., and Zhao, Y. (2006). Auxin biosynthesis by the YUCCA flavin monooxygenases controls the formation of floral organs and vascular tissues in Arabidopsis. *Genes Dev.* 20, 1790–1799.
- Cheng, Y., Dai, X., and Zhao, Y. (2007). Auxin Synthesized by the YUCCA Flavins Monooxygenases Is Essential for Embryogenesis and Leaf Formation in Arabidopsis. *Plant Cell* 19, 2430–2439.
- Cooke, T.J., Poli, D., Sztein, A.E., and Cohen, J.D. (2002). Evolutionary patterns in auxin action. *Plant Mol. Biol.* 49, 319–338.
- Dillman, A.R., Minor, P.J., and Sternberg, P.W. (2013). Origin and Evolution of Dishevelled. *G3 Genes|Genomes|Genetics* 3, 251–262.
- van Dop, M., Fiedler, M., Mutte, S., de Keijzer, J., Olijslager, L., Albrecht, C., Liao, C.-Y., Janson, M.E., Bienz, M., and Weijers, D. (2020). A conserved biochemical paradigm underlies cell polarity across multicellular kingdoms. *Cell.* in press.
- Eklund, D.M., Ishizaki, K., Flores-Sandoval, E., Kikuchi, S., Takebayashi, Y., Tsukamoto, S., Hirakawa, Y., Nonomura, M., Kato, H., Kouno, M., et al. (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.
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152.
- Gao, Y., Zhang, Y., Zhang, D., Dai, X., Estelle, M., and Zhao, Y. (2015). Auxin binding protein 1 (ABP1) is not required for either auxin signaling or Arabidopsis development. *Proc. Natl. Acad. Sci.* 112, 2275–2280.
- Katsir, L., Schilmiller, A.L., Staswick, P.E., He, S.Y., and Howe, G.A. (2008). COI1 is a critical component of a receptor for jasmonate and the bacterial virulence factor coronatine. *Proc. Natl. Acad. Sci.* 105, 7100–7105.
- Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMET-SP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* 12, e1001889.
- Kuettner, E.B., Hilgenfeld, R., and Weiss, M.S. (2002). The Active Principle of Garlic at Atomic Resolution. *J. Biol. Chem.* 277, 46402–46407.
- Ljung, K. (2013). Auxin metabolism and homeostasis during plant development. *Development* 140, 943–950.
- Lu, K.-J., van 't Wout Hoffland, N., Mor, E., Mutte, S., Abrahams, P., Kato, H., Vandepoele, K., Weijers, D., and De Rybel, B. (2020). Evolution of vascular plants through redeployment of ancient developmental regulators. *Proc. Natl. Acad. Sci.* in press.
- Ludwig-Müller, J. (2011). Auxin conjugates: their role for plant development and in the evolution of land plants. *J. Exp. Bot.* 62, 1757–1773.
- Mashiguchi, K., Tanaka, K., Sakai, T., Sugawara, S., Kawaide, H., Natsume, M., Hanada, A., Yaeno, T., Shirasu, K., Yao, H., et al. (2011). The main auxin biosynthesis pathway in Arabidopsis. *Proc. Natl. Acad. Sci.* 108, 18512–18517.
- Mutte, S.K., Kato, H., Rothfels, C., Melkonian, M., Wong, G.K.-S., and Weijers, D. (2018). Origin and evolution of the nuclear auxin response system. *Elife* 7, e33399.
- Nobuta, K., Okrent, R.A., Stoutemyer, M., Rodibaugh, N., Kempema, L., Wildermuth, M.C., and Innes, R.W. (2007). The GH3 Acyl Adenylase Family Member PBS3 Regulates Salicylic Acid-Dependent Defense Responses in Arabidopsis. *Plant Physiol.* 144, 1144–1156.
- Nonhebel, H.M. (2015). Tryptophan-Independent Indole-3-Acetic Acid Synthesis: Critical Evaluation of the Evidence. *Plant Physiol.* 169, 1001–1005.
- Okrent, R.A., Brooks, M.D., and Wildermuth, M.C. (2009). Arabidopsis GH3.12 (PBS3) Conjugates Amino Acids to 4-Substituted Benzoates and Is Inhibited by Salicylate. *J. Biol. Chem.* 284, 9742–9754.
- Östlin, A., Kowalczyk, M., Bhalerao, R.P., and Sandberg, G. (1998). Metabolism of Indole-3-Acetic Acid in Arabidopsis.

Plant Physiol. 118, 285–296.

- Paponov, I.A., Dindas, J., Król, E., Friz, T., Budnyk, V., Teale, W., Paponov, M., Hedrich, R., and Palme, K. (2019). Auxin-Induced Plasma Membrane Depolarization Is Regulated by Auxin Transport and Not by AUXIN BINDING PROTEIN1. *Front. Plant Sci.* 9.
- Robert, S., Kleine-Vehn, J., Barbez, E., Sauer, M., Paciorek, T., Baster, P., Vanneste, S., Zhang, J., Simon, S., Čovanová, M., et al. (2010). ABP1 Mediates Auxin Inhibition of Clathrin-Dependent Endocytosis in Arabidopsis. *Cell* 143, 111–121.
- Romani, F. (2017). Origin of TAA Genes in Charophytes: New Insights into the Controversy over the Origin of Auxin Biosynthesis. *Front. Plant Sci.* 8, 2016–2018.
- De Rybel, B., Möller, B., Yoshida, S., Grabowicz, I., Barbier de Reuille, P., Boeren, S., Smith, R.S., Borst, J.W., and Weijers, D. (2013). A bHLH Complex Controls Embryonic Vascular Tissue Establishment and Indeterminate Growth in Arabidopsis. *Dev. Cell* 24, 426–437.
- Staswick, P.E., Tiryaki, I., and Rowe, M.L. (2002). Jasmonate Response Locus JAR1 and Several Related Arabidopsis Genes Encode Enzymes of the Firefly Luciferase Superfamily That Show Activity on Jasmonic, Salicylic, and Indole-3-Acetic Acids in an Assay for Adenylation. *Plant Cell* 14, 1405–1415.
- Staswick, P.E., Serban, B., Rowe, M., Tiryaki, I., Maldonado, M.T., Maldonado, M.C., and Suza, W. (2005). Characterization of an Arabidopsis Enzyme Family That Conjugates Amino Acids to Indole-3-Acetic Acid. *Plant Cell* 17, 616–627.
- Stepanova, A.N., Robertson-Hoyt, J., Yun, J., Benavente, L.M., Xie, D.-Y., Doležal, K., Schlereth, A., Jürgens, G., and Alonso, J.M. (2008). TAA1-Mediated Auxin Biosynthesis Is Essential for Hormone Crosstalk and Plant Development. *Cell* 133, 177–191.
- Stepanova, A.N., Yun, J., Robles, L.M., Novak, O., He, W., Guo, H., Ljung, K., and Alonso, J.M. (2011). The Arabidopsis YUCCA1 Flavin Monooxygenase Functions in the Indole-3-Pyruvic Acid Branch of Auxin Biosynthesis. *Plant Cell* 23, 3961–3973.
- Turnaev, I.I., Gunbin, K. V., and Afonnikov, D.A. (2015). Plant auxin biosynthesis did not originate in charophytes. *Trends Plant Sci.* 20, 463–465.
- Wang, B., Chu, J., Yu, T., Xu, Q., Sun, X., Yuan, J., Xiong, G., Wang, G., Wang, Y., and Li, J. (2015). Tryptophan-independent auxin biosynthesis contributes to early embryogenesis in Arabidopsis. *Proc. Natl. Acad. Sci.* 112, 4821–4826.
- Wang, C., Li, S.-S., and Han, G.-Z. (2016). Commentary: Plant Auxin Biosynthesis Did Not Originate in Charophytes. *Front. Plant Sci.* 7, 158.
- Weijers, D., and Wagner, D. (2016). Transcriptional Responses to the Auxin Hormone. *Annu. Rev. Plant Biol.* 67, 539–574.
- Westfall, C.S., Herrmann, J., Chen, Q., Wang, S., and Jez, J.M. (2010). Modulating plant hormones by enzyme action. *Plant Signal. Behav.* 5, 1607–1612.
- Westfall, C.S., Zubieta, C., Herrmann, J., Kapp, U., Nanao, M.H., and Jez, J.M. (2012). Structural Basis for Preceptor Modulation of Plant Hormones by GH3 Proteins. *Science* 336, 1708–1711.
- Westfall, C.S., Shero, A.M., Zubieta, C., Alvarez, S., Schraft, E., Marcellin, R., Ramirez, L., and Jez, J.M. (2016). Arabidopsis thaliana GH3.5 acyl acid amido synthetase mediates metabolic crosstalk in auxin and salicylic acid homeostasis. *Proc. Natl. Acad. Sci.* 113, 13917–13922.
- Woo, E.-J. (2002). Crystal structure of auxin-binding protein 1 in complex with auxin. *EMBO J.* 21, 2877–2885.
- Xu, T., Dai, N., Chen, J., Nagawa, S., Cao, M., Li, H., Zhou, Z., Chen, X., De Rycke, R., Rakusova, H., et al. (2014). Cell Surface ABP1-TMK Auxin-Sensing Complex Activates ROP GTPase Signaling. *Science* 343, 1025–1028.
- Yoshida, S., van der Schuren, A., van Dop, M., van Galen, L., Saiga, S., Adibi, M., Möller, B., ten Hove, C.A., Marhavy, P., Smith, R., et al. (2019). A SOSEKI-based coordinate system interprets global polarity cues in Arabidopsis. *Nat. Plants* 5, 160–166.
- Zhang, L., Zhang, F., Melotto, M., Yao, J., and He, S.Y. (2017). Jasmonate signaling and manipulation by pathogens and insects. *J. Exp. Bot.* 68, 478.
- Zhao, Y. (2012). Auxin Biosynthesis: A Simple Two-Step Pathway Converts Tryptophan to Indole-3-Acetic Acid in Plants. *Mol. Plant* 5, 334–338.
- Zhao, Y. (2014). Auxin Biosynthesis. *Arab. B.* 12, e0173.
- Zhao, Y. (2018). Essential Roles of Local Auxin Biosynthesis in Plant Development and in Adaptation to Environmental Changes. *Annu. Rev. Plant Biol.* 69, 417–435.



Origin and evolution of the nuclear auxin response system

Sumanth K. Mutte^{1,*}, Hirotaka Kato^{1,*}, Carl Rothfels², Michael Melkonian³, Gane Ka-Shu Wong^{4,5,6}, and Dolf Weijers¹

¹Laboratory of Biochemistry, Wageningen University, 6708WE Wageningen, the Netherlands

²Department of Integrative Biology, University of California, Berkeley, CA, United States of America

³Botanical Institute, Cologne Biocenter, University of Cologne, D50674 Cologne, Germany

⁴Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

⁵Department of Medicine, University of Alberta, Edmonton, Alberta, Canada

⁶BGI-Shenzhen, Bei Shan Industrial Zone, Yantian District, Shenzhen, China

This chapter has been published as:

Mutte, S.K.*, Kato, H.*, Rothfels, C., Melkonian, M., Wong, G.K.-S., and Weijers, D. (2018). Origin and evolution of the nuclear auxin response system. *Elife* 7, e33399.

* These authors contributed equally



The small signalling molecule auxin controls numerous developmental processes in land plants, acting mostly by regulating gene expression. Auxin response proteins are represented by large families of diverse functions, but neither their origin nor their evolution is understood. Here we use a deep phylogenomics approach to reconstruct both the origin and the evolutionary trajectory of all nuclear auxin response protein families. We found that, while all subdomains are ancient, a complete auxin response mechanism is limited to land plants. Functional phylogenomics predicts defined steps in the evolution of response system properties, and comparative transcriptomics across six ancient lineages revealed how these innovations shaped a sophisticated response mechanism. Genetic analysis in a basal land plant revealed unexpected contributions of ancient non-canonical proteins in auxin response as well as auxin-unrelated function of core transcription factors. Our study provides a functional evolutionary framework for understanding diverse functions of the auxin signal.

Introduction

Auxins are a group of structurally related chemical compounds that control a multitude of growth and developmental processes in plants. The most common, naturally occurring auxin is Indole-3-acetic acid (IAA), but synthetic analogs such as 2,4-dichlorophenoxy acetic acid (2,4-D) have largely overlapping biological activities (Woodward and Bartel, 2005). While auxins have been shown to trigger rapid cellular events such as membrane hyperpolarization (Bates and Goldsmith, 1983; Etherton, 1970), calcium influx (Monshausen et al., 2011; Schenck et al., 2010), and changes in endocytosis (Paciorek et al., 2005; Robert et al., 2010), its activity in controlling growth and development appear to be mainly mediated by changes in gene expression via a nuclear auxin pathway (NAP). Perturbation of this gene regulatory pathway interferes with most, if not all, developmental responses (Weijers and Wagner, 2016). Indeed, in the moss *Physcomitrella patens*, it was shown that a complete knock-out mutant of this pathway does not show any transcriptional response to auxin (Lavy et al., 2016). The NAP encompasses three dedicated protein families (Fig. 1A, B). Various auxins, including IAA and 2,4-D are perceived by a co-receptor complex consisting of TRANSPORT INHIBITOR RESPONSE 1/AUXIN SIGNALING F-BOX (TIR1/AFB) and AUXIN/INDOLE-3-ACETIC ACID (Aux/IAA) proteins (Dharmasiri et al., 2005; Kepinski and Leyser, 2005; Tan et al., 2007). Subsequent ubiquitination of the Aux/IAA proteins causes their degradation in the 26S proteasome (Gray et al., 2001). When not degraded, Aux/IAA proteins bind to and inhibit DNA-binding transcription factors, the AUXIN RESPONSE FACTORS (ARF) (Kim et al., 1997). Thus, auxin de-represses ARFs, allowing these to activate or repress their direct target genes (Ulmasov et al., 1999).

A central question in plant biology is how this simple transcriptional system with only three dedicated components can generate a multitude of local auxin responses to support various developmental functions. In flowering plants such as *Arabidopsis thaliana*, it is likely that the size of TIR1/AFB (six members), Aux/IAA (29 members) and ARF (23 members) gene families allows combinatorial assembly of distinct, local auxin response pathways. Given that diversity in auxin responses follows from diversification in its response proteins, it is still unclear how NAP complexity evolved from simpler ancestral states. Furthermore, while intuitive, a key question is whether increased NAP complexity indeed enabled more complex and diverse auxin responses during plant evolution. A third important question is where, when, and from what precursors the NAP originated.

Eukaryotic photosynthetic organisms diverged into three groups, Glaucophyta, Rhodophyta (red algae), and Viridiplantae more than 1.5 billion years ago (Yoon et al., 2004). Viridiplantae are further classified into chlorophyte algae and streptophytes, which include charophyte algae and land plants. Bryophytes represent the earliest diverging land plants and consist of three groups: hornworts, liverworts and mosses. After the split from bryophytes, ancestral vascular plants changed their life cycle from haploid-dominant to diploid-dominant and established a vascular system and root architecture, forming the group of lycophytes and euphyllophytes (ferns, gymnosperms and angiosperms).

The presence of a functional NAP with reduced genetic redundancy has been reported in model bryophytes (Flores-Sandoval et al., 2015; Kato et al., 2015; Prigge et al., 2010; Rensing et al., 2008), whereas the presence of endogenous auxin is also reported in wide range of algal species (Žižková et al., 2017). Thus, a prediction is that the auxin response system may predate land plants, and that complexity evolved after the divergence of ancestral vascular plants from bryophytes. A key challenge is to identify the origin of the NAP system, as well as to reconstruct the steps in the evolution of its complexity. However, only little genome data are currently available from non-flowering land plants (Rensing, 2017), which makes such inferences extremely challenging. In addition, studies using only selected model species bear the risk of generalizing observations from non-representative genomes, due to species-specific gene-duplication, -loss, and -diversification. Therefore, it is necessary to analyse multiple species to understand evolutionary trends.

Here we describe a deep phylogenomic analysis of NAP components using a large transcriptome dataset with more than 1,000 plant species including many algae. This extensive dataset enabled us to reconstruct the ancestral states of auxin response gene families at key nodes in plant evolution. We infer plausible origins and evolutionary patterns for each auxin response gene family and predict auxin response properties at evolutionary nodes. Using comparative RNA-Seq of six species, we tested and extended these predictions. Finally, we used a genetic strategy in a bryophyte to demonstrate surprising non-contributions of an ancient ARF class as well as contribution of deeply conserved non-canonical NAP components to auxin signalling. Our work provides a deep view into early steps in the origin, evolution and design principles of the multi-functional auxin response system.

Results

A phylogenomic strategy for reconstructing ancestral states

To reconstruct origin and early diversification in auxin response gene families, we designed a strategy (Chapter 2) that uses a large transcriptome dataset (OneKP) including multiple species for each major branch in plant species phylogeny (Matasci et al., 2014). The depth and quality of each individual RNA-Seq-derived transcriptome is limited and a further caveat of transcriptome-based gene identifications is that the number of genes may be underestimated if a gene is not expressed under the sampling conditions or in the sampled tissue. However, the availability of transcriptomes from multiple tissue samples of multiple related species, should allow deduction of the ancestral state that defines the gene complement at each evolutionary node. It should be stressed that this number represents the ancestral state at a given node, and species-specific gene duplications and gene losses will have modified the gene complement in individual species. Given our focus on early events in nuclear auxin response evolution, we have used all available transcriptomes of red algae, green algae, bryophytes, lycophytes, ferns, and gymnosperms from the OneKP dataset (Supplementary file 1). We also included all available angiosperm species in the Chloranthales, Magnoliids and ANA grade, as well as several species in both monocots and dicots (Supplementary file 1). For reference and quality control purposes, we included genome-

based sequences from well annotated model species.

Origin of nuclear auxin response components

Each of the three auxin response protein types (ARFs, Aux/IAAs and TIR1/AFBs) are multi-domain proteins and we initially focused on the origin of these proteins. Therefore, we asked where domains, or parts thereof, were found, and at what node the multi-domain proteins first appear. ARF proteins carry an N-terminal DNA-binding domain (DBD) which consists of a composite dimerization domain (DD; made up of two separate subdomains [DD1 and DD2] that fold into a single unit), a B3-type DNA-interaction domain, and an ancillary domain (AD) of unknown function (Fig. 1C; Boer et al., 2014). In land plants, the DD and AD are only found in the ARF family. The C-terminal Phox and Bem1 (PB1) domain is shared among ARF and Aux/IAA proteins and mediates homo- and hetero-oligomerization (Korasick et al., 2014; Nanao et al., 2014). Finally, ARFs contain a less well-defined Middle Region (MR) separating the PB1 and DBD (Fig. 1C).

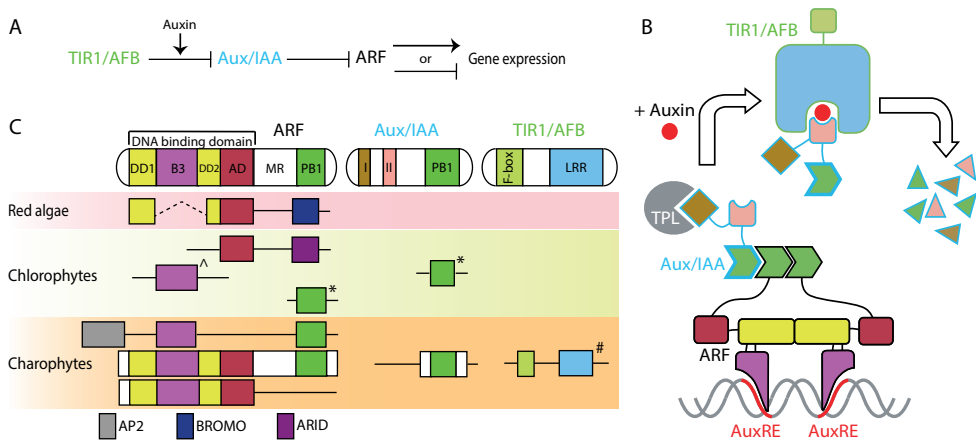


Figure 1: Proteins in nuclear auxin pathway; mechanism and origin of the domains. (A, B) Scheme of NAP in land plants. In the absence of auxin, Aux/IAA inhibit ARF via their PB1 domains, and by recruiting the TPL co-repressor. Auxin stabilizes the interaction between Aux/IAA and TIR1/AFB, followed by proteasome-mediated degradation of Aux/IAA. (C) Domain structure of NAP components in land plants and presence of each domain in algae, as recovered in transcriptomes. (Sub-)domains are indicated by colors, that match those in (B). \wedge : basal to all B3-type transcription factors in land plants, *: difficult to assign to ARF or Aux/IAA family; #: forming basal clade to both TIR1/AFB and COI1 in land plants.

In red algae, we found proteins containing an N-terminal portion of DD1, DD2, and AD, lacking a B3 or PB1 domain, but instead flanked by a C-terminal bromodomain (BROMO; InterPro ID: IPR001487; Fig. 1C). The DD1 and DD2 motifs in red algae are spaced by 20–30 conserved amino acids, which is much shorter than the B3 domain (~120 amino acids; Supplementary file 2). In chlorophytes, we found a protein with only AD, flanked by a DNA-binding AT-rich interaction domain (ARID; InterPro ID: IPR001606; Fig. 1C). Furthermore, we found separate proteins that either represented a B3 or a PB1 domain (Fig. 1C). Thus, all ARF

subdomains had been established before the split of the streptophytes, but not combined in a single protein. In contrast, we discovered full-length ARF-like proteins containing a DBD with a B3 domain inserted between DD and AD in charophytes (Fig. 1C and Fig. S1). Land plant ARFs can be grouped into three classes, A, B and C (Finet et al., 2013). Based on transactivation assays, class A- and B-ARFs are classified as transcriptional activators and repressors, respectively (Kato et al., 2015; Ulmasov et al., 1999). Class C-ARFs are generally recognized as transcriptional repressors based on the amino acid composition of MR, but this has not yet been fully supported by experimental evidence (Kato et al., 2017b). Phylogenetic analysis revealed that the ARF-like proteins in charophytes fall in two sister clades and likely represent separate precursors of class C-ARFs (proto-C-ARFs) and A/B-ARFs (proto-A/B-ARFs) of land plants (Fig. 2 and Fig. S1). Interestingly, we found the PB1 domain only in proto-C-ARFs, which could however be due to sparse sampling in some charophyte lineages (Fig. S1).

To understand if the proto-ARFs share conserved, functionally important residues, we generated homology models based on available DBD crystal structures of *A. thaliana* ARF1 and ARF5 (Boer et al., 2014). As no class C-ARF structure is known, we first modelled the *A. thaliana* ARF10 DBD to compare with proto-C-ARFs. Next, homology models for proto-ARFs in *Spirogyra pratensis* (SpARF; proto-C-ARF) and *Mesotaenium caldarium* (McARF; proto-A/B-ARF) were generated. We also included all three ARFs of the bryophyte *M. polymorpha* (MpARF1-3) representing each major class, and compared all models to *A. thaliana* ARF structures. This analysis revealed that all proto-ARFs likely share a conserved structural topology (Fig. 3A). Strikingly, all DNA-binding residues follow the spatial restraints needed for DNA binding in all ARFs tested, suggesting a conserved mode of DNA binding. On the other hand, dimerization residues are conserved only in the (proto-)A/B-ARFs (McARF, MpARF1, and MpARF2) but not in the (proto-)C-ARFs (SpARF, MpARF3, and ARF10) (Fig. 3A). These results clearly demonstrate that canonical ARF proteins were established and differentiated into two classes in charophyte algae.

In addition to the proteins with canonical ARF-like structure, we found a group of charophyte proteins consisting of an AP2 DNA binding domain along with B3 and PB1 domains (Fig. 1C). Land plants also have a protein family containing AP2 domain in their N-terminus, followed by a B3 domain. These proteins are called RELATED TO ABI3 AND VP1 (RAV). Interestingly, land plant RAV proteins do not have a PB1 domain and it is known that the B3 domain of RAV and ARF binds different DNA sequences (Boer et al., 2014; Matias-Hernandez et al., 2014). The B3 domain of RAV-like proteins in charophytes is much more similar to RAV than to ARF proteins in land plants (Fig. S2). Phylogenetic analysis showed that the RAV-like proteins of charophytes position along with RAV family in land plants (Fig. 2B and Fig. S1). Thus, we classify these proteins as proto-RAV. In the charophyte green algae, the two classes of proto-ARFs and proto-RAVs are found in various combinations in each species (Fig. 2A). While sequencing depth may be insufficient to detect all proto-ARFs and proto-RAVs, there does not appear to be a conserved pattern in the order of appearance and retention of these genes.

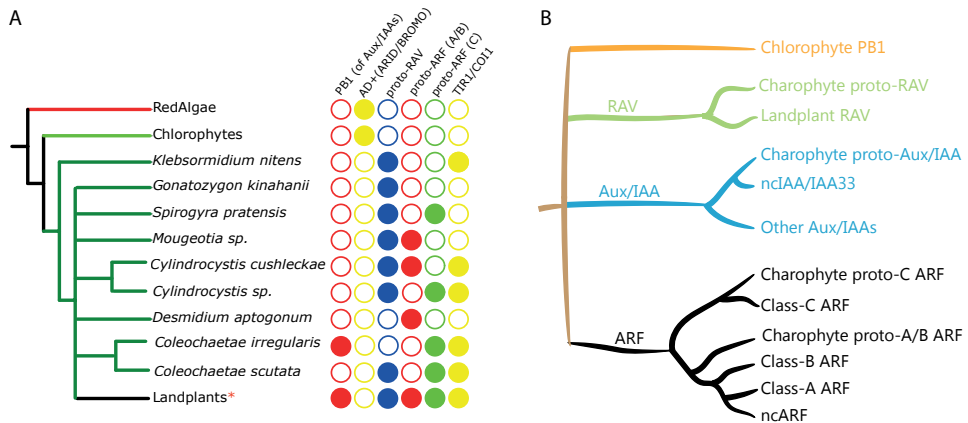


Figure 2: Distribution of auxin signalling proteins precursors in algal lineages. (A) Occurrence of NAP components in red algae, chlorophytes, and charophytes. Empty circles and filled circles indicate the absence and presence of that particular component, respectively. *: Land plants have defined three classes of ARFs, RAV without PB1, and separate TIR1/AFB and COI1 receptors. (B) Schematic illustration of the phylogenetic arrangement of RAV1, Aux/IAA and ARFs based on the DBD tree and PB1 tree. Note that only branches with strong bootstrap support are shown.

We next considered the origin of the Aux/IAA proteins. These proteins contain two functional small domains in addition to a C-terminal PB1 domain (Fig. 1B, C). The N-terminal domain I recruits the TOPLESS (TPL) transcriptional co-repressor (Szemenyei et al., 2008). Domain II mediates the auxin-dependent interaction with TIR1/AFB and thus acts as a degron (Dharmasiri et al., 2005; Gray et al., 2001; Kepinski and Leyser, 2005). Because domain I and II are too small for reliable BLAST searches, we used the PB1 domain to identify potential family members. No PB1-containing proteins were identified in red algae, while we found proteins with a PB1 domain but no DBD in chlorophytes (Fig. 1C). Phylogenetic analysis based on the PB1 domain indicated these are neither closely related to RAV, nor to Aux/IAA and ARF families (Fig. 2B and Fig. S3). PB1 domain-containing proteins that lack a DBD were also found in many of the charophyte algae (Fig. 1C, Fig. 3B and Fig. S3). Most of them were placed along with proto-RAV in phylogenetic tree, but the sequences from *Coleochaetae irregularis* were placed along with the Aux/IAA in land plants that is separate from the PB1 of both ARFs and proto-RAV proteins (Fig. 2B, Fig. 3B and Fig. S3). Even though the N-terminal part of the PB1 domain is not as conserved as the C-terminal part, several critical residues were found to be conserved in Aux/IAA-like sequences (Fig. 3B). These results indicate that the PB1 domain of land plant ARFs and Aux/IAAs had separate precursors in charophytes. We could, however, not detect domain I or II in Aux/IAA-like genes of charophyte algae, even when scrutinizing individual sequences. We thus conclude that Aux/IAA proteins with all three functional domains are limited to land plants.

Finally, we explored the origin of the TIR1/AFB auxin co-receptor that consist of an N-terminal F-box domain that anchors the protein to the other subunits in the SCF E3 ubiquitin ligase complex, and a C-terminal leucine-rich repeat (LRR) domain that contains the auxin binding pocket. Auxin acts as a molecular glue to stabilize the interaction between TIR1/

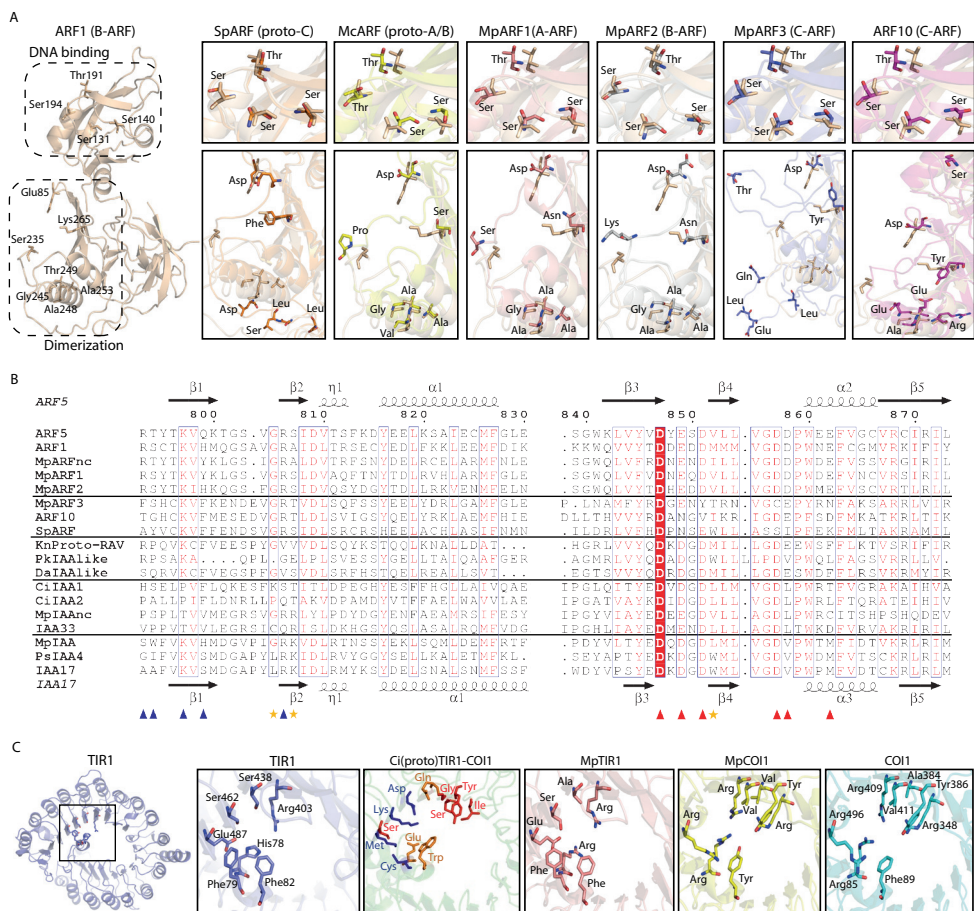


Figure 3: Homology models of ancestral ARF, Aux/IAA and TIR1/AFB proteins. (A) Homology models for ARF DBDs. The crystal structure of *Arabidopsis thaliana* ARF1-DBD is shown on the left with important residues for DNA binding (top) and dimerization (bottom). Homology models for (proto-) ARFs are overlaid on *A. thaliana* ARF1 in right panels (brown). (B) Alignment of PB1 domain of (proto-) ARF, Aux/IAA and proto-RAV proteins. Numbering is based on the ARF5 protein of *A. thaliana*. Arrows and helices indicate β -sheets and α -helices in ARF5 and IAA17 of *A. thaliana*, respectively. Blue and red triangles indicate positive (+) and negative (-) faces, respectively. Golden asterisks represent the residues of polar interactions. (C) Homology models for TIR1/AFB and COI1 proteins. Left panel shows crystal structure of *A. thaliana* TIR1 from top view. Auxin binding pocket of TIR1/AFB and jasmonate binding pocket of COI1 are shown in right panels. Hormone binding residues are indicated as stick model in TIR1 and COI1 of land plants. Blue, red or orange residues in the model for the *Coleochaete irregularis* protein indicate the residues aligned with hormone binding residues of TIR1, COI1 or both, respectively. Ci: *Coleochaete irregularis*, Da: *Desmidiium aptogonum*, Kn: *Klebsormidium nitens*, Mc: *Mesotaenium caldarium*, Mp: *Marchantia polymorpha*, Pk: *Parachlorella kessleri*, Ps: *Pisum sativum*, Sp: *Spirogyra pratensis*.

AFB and Aux/IAA proteins (Tan et al., 2007). The closest homolog of the TIR1/AFB proteins in *A. thaliana* is CORONATINE INSENSITIVE 1 (COI1), which functions as a receptor of the jasmonic acid (JA) phytohormone (Katsir et al., 2008). In our homology search, we could not identify any proteins showing homology to either TIR1/AFB or COI1 in red algae and chlorophytes (Fig. 1C and Fig. 2A). We did find many proteins showing homology to TIR1/

AFB and COI1 in the transcriptomes of charophyte algae (Fig. 1C and Fig. 2A). Phylogenetic analysis indicated that some of these proteins form a sister group to both TIR1/AFB and COI1 in land plants and none of the charophyte proteins are specifically grouped into either TIR1/AFB or COI1 clades (Fig. 4, Fig. S5 and Fig. S6), suggesting that charophytes had an ancestor that gave rise to both auxin and JA receptors.

To infer whether the TIR1/AFB/COI1-like proteins of charophytes function as receptors for auxin or JA, we generated homology models of the TIR1/AFB/COI1-like protein from *C. irregularis* and the bryophyte *M. polymorpha* MpTIR1 and MpCOI1, using the *A. thaliana* TIR1 and COI1 crystal structures (Sheard et al., 2010; Tan et al., 2007) as templates for modelling. Even though the secondary structure of the *C. irregularis* protein was highly similar to that of land plant TIR1 and COI1 (Supplementary file 2), at the level of amino acid sequence, the protein did not resemble either TIR1/AFB or COI1. Out of 40 residues conserved in either TIR1/AFB or COI1, only 7 and 11 residues are identical to TIR1/AFB and COI1, respectively (Supplementary file 2; black stars). Notably, most of the hormone-contacting residues (11 out of 12) are different from both TIR1/AFB and COI1 (Fig. 3C and Supplementary file 2). These results suggest that the charophyte TIR1/AFB/COI1 precursor may not act as an auxin or JA receptor, and we conclude that dedicated receptors for auxin and JA were established only in land plants. Taken together, our analyses suggest that the components of NAP were established in the common ancestor of land plants by combining pre-existing components and that the system evolved to regulate pre-existing transcription factors.

Evolution of complexity in the nuclear auxin response system

All three gene families have evolved to considerable size and diversity in angiosperms, and this diversity is thought to underlie multifunctionality of auxin as a hormone. We next aimed to reconstruct the evolutionary history of auxin response components across all land plant lineages. Consistent with previous descriptions (Finet et al., 2013), our phylogenetic analysis showed that all land plant ARFs are divided into three phylogenetic lineages (Fig. 4 and Fig. S1). Within the class-C lineage, we did not find any duplications in the ancestors of non-angiosperm species. The split that generated *A. thaliana* ARF10/16 and ARF17 likely occurred early in angiosperm evolution, while the PB1 domain was lost in the ARF17 group (Fig. 4 and Fig. S1). The class A-ARF is represented by a single copy in bryophytes and lycophytes. We found that a subset of genes lacking the DBD diverged from class A-ARFs in early land plants, is missing in hornworts and has been retained in liverworts, mosses and lycophytes (non-canonical ARF, ncARF; Fig. 3B, Fig. 4 and Fig. S3). A further gene duplication event in the ancestor of euphyllophytes gave rise to two class-A sub-families corresponding to *A. thaliana* ARF5/7/19 and ARF6/8, respectively. In the ancestor of seed plants, a gene duplication caused differentiation between the *A. thaliana* ARF5 and ARF7/19 subfamilies (Fig. 4 and Fig. S1). Finally, two gene duplication events in the ancestral angiosperms led to ARF6 and ARF8 and to a paralog of ARF7/19, which was lost in *A. thaliana* (Fig. 4 and Fig. S1).

Class B-ARFs are represented by a single gene in the ancestor of liverworts, mosses, lycophytes, and ferns. However, no hornwort species appears to contain class B-ARFs (Fig. 4 and Fig. S1). Gene duplications in the ancestral gymnosperms gave rise to three class B-ARF copies, one representing *A. thaliana* ARF3/4, another leading to *A. thaliana* ARF2 and the third generating the remainder of the class B-ARFs in *A. thaliana* (Fig. 4 and Fig. S1). Notably, the reported lack of the PB1 domain in ARF3 (Finet et al., 2013) is an independent loss in the common ancestor of monocots and eudicots (Fig. S1).

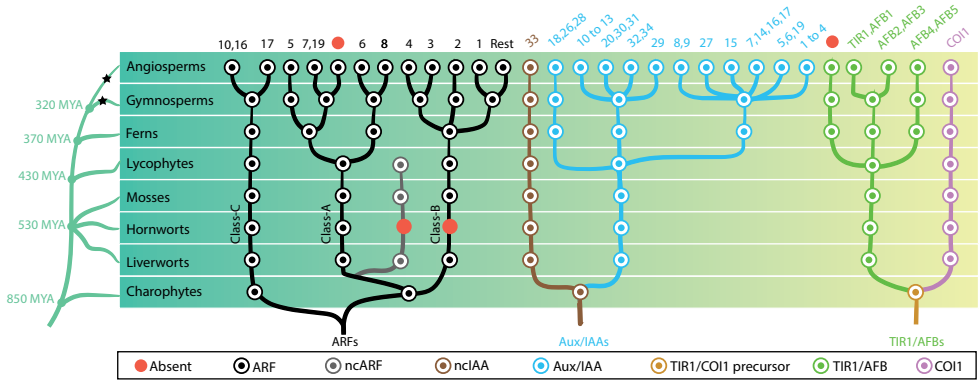


Figure 4: Reconstruction of ancestral state of NAP components in plant evolution. Schematic summary shows the ancestral copy number and phylogenetic relationship of each protein family in all major plant lineages. Each circle is colored according to protein type as indicated in the box. In the top row, numbers indicate which genes of *Arabidopsis thaliana* belong to each subfamily and red circles indicates missing subfamilies in *A. thaliana*. Note that only branches with strong bootstrap support are shown.

Our data indicated that an ancestral Aux/IAA gene lacking domain I and II had been established during the evolution of charophytes, while “true” Aux/IAs with all functional domains are found only in land plants (Fig. 1C). In addition to one copy of “true” Aux/IAA, we found another set of deeply conserved non-canonical Aux/IAA-like sequences that lack the domain I and II (non-canonical Aux/IAA, ncIAA; Fig. 2B, Fig. 4, Fig. S3 and Fig. S4). Strikingly, while the Aux/IAs have diversified through gene duplications, the ncIAA is found only in a single copy in all evolutionary nodes examined here, and is represented by IAA33 in *A. thaliana*. In the ancestor of euphyllophytes, gene duplication events gave rise to three Aux/IAs, which were retained in the ancestral seed plants (Fig. 4 and Fig. S4). Common ancestor of angiosperms has eleven Aux/IAA proteins, which is more than triple the number found in gymnosperms (Fig. 4 and Fig. S4). Finally, in addition to the ancient ncIAA generated in a first duplication event, several independent events later generated non-canonical family members lacking domains. For example, the lack of domain II in IAA20, IAA30, IAA31, IAA32, and IAA34 of *A. thaliana* appears to be an independent loss in their respective lineages in the core angiosperms (Fig. S4).

Our data indicated that ancestral charophyte green algae had one common ancestor for both auxin (TIR1/AFB) and JA (COI1) F-box co-receptors, and following duplication in the ancestor of all land plants, developed into two independent receptors (Fig. 4 and Fig. S5). The common ancestor of bryophytes and lycophytes had a single orthologue of *A. thaliana* TIR1/

AFB. Gene duplication events in the ancestor of euphyllophytes gave rise to three subgroups; one leading to TIR1/AFB1-3, one leading to AFB4/5 and another which is widely present in many species including the angiosperms, but has been lost in some monocots and dicots including *A. thaliana* (Fig. 4, Fig. S5 and Fig. S6).

Thus, our analysis of the patterns of diversification in the ARE, Aux/IAA and TIR1/AFB families identifies the auxin response complement at each evolutionary node, and in addition reveals deeply conserved non-canonical family members. Notably, many changes occurred in the composition of NAP from the common ancestor of lycophytes to euphyllophytes, which may have led to complex auxin response.

Multi-species comparative transcriptome analysis reveals evolution of response complexity

The complements of auxin response components identified from phylogenomic analysis allow for clear predictions of which species possess a functional transcriptional auxin response system. Based on our predictions, only land plants should be able to respond. In addition, it is intuitive that the number of components in auxin response will relate to the complexity of response, but as yet there is no experimental basis for such relationship. To experimentally address the competence of species to respond to auxin, and to explore the relationship between auxin response components and the qualitative and quantitative aspects of auxin response, we performed comparative transcriptome analysis. We selected six species that belong to different ancient lineages and that each have a different complement of auxin response components (Fig. 5A). We used the charophyte algae *Klebsormidium nitens* and *Spirogyra pratensis*, the hornwort *Anthoceros agrestis*, the liverwort *Marchantia polymorpha*, the moss *Physcomitrella patens*, and the fern *Ceratopteris richardii*. To detect only early transcriptional responses, we treated plants with auxin for 1h, and performed RNA-Seq followed by de novo transcriptome assembly and differential gene expression analysis. To avoid inactivation of the natural auxin IAA by conjugation or transport, we treated with 10 μ M of the synthetic auxin 2,4-dichlorophenoxyacetic acid (2,4-D). This compound was shown to behave like IAA in the context of the NAP (Tan et al., 2007).

Importantly, 68–90% of the differentially expressed genes (DEG) from de novo assemblies in *K. nitens*, *M. polymorpha* and *P. patens* matched with genome-based differential gene expression performed in parallel (Fig. S7), thus validating our approach. Transcriptome analysis after prolonged auxin treatment in *P. patens* had identified a large set of auxin-responsive genes (Lavy et al., 2016). Indeed, we found 105 and 1090 genes to be auxin-regulated in *M. polymorpha* and *P. patens*, respectively (Fig. 5A). Likewise, we found 159 and 413 genes to be auxin-regulated in *A. agrestis* and *C. richardii* (Fig. 5A). Unexpectedly, despite lacking Aux/IAA and dedicated TIR1/AFB genes, both charophyte algae showed a strong transcriptional response to 2,4-D treatment. A total of 1094 and 1681 genes were differentially expressed in *K. nitens* and *S. pratensis*, respectively (Fig. 5A). Thus, there is a clear transcriptional response to 1 hour of 2,4-D treatment in all species analyzed, yet the number of genes is different, with an exceptionally large number of responsive genes in charophytes. We next determined if the number of DEG

correlates with gene number in each transcriptome assembly (Fig. S8), and found that differences in DEG among species cannot be explained by total gene number.

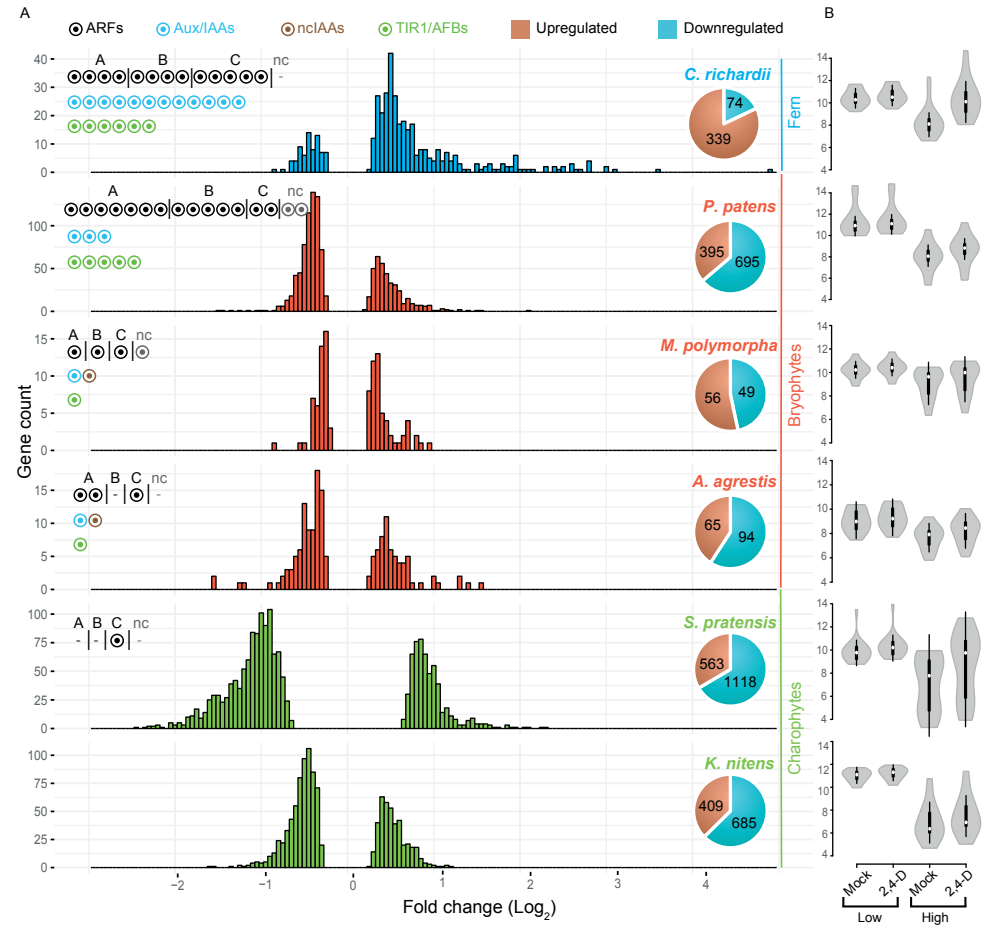


Figure 5: Auxin-dependent gene regulation across basal plant species. (A) Histograms represent the distribution of log₂ fold change among differentially expressed genes on X-axis ($\text{Padj} < 0.01$). Y-axis indicates the number of genes in each log₂ fold-change bin. Pie charts indicate the total number of up- and down-regulated genes in each species. Circles in the top left of each graph indicate the number of NAP components. **(B)** Violin plots of log₂ normalized expression values (DESeq2-based; y-axis) of 20 least auxin-activated (Low) and 20 top-most auxin-upregulated (High) genes in each six species. White dot indicated the median expression value.

We next addressed whether there were differences in the characteristics of regulation. Both charophyte species showed a high percentage of gene repression. Only 37% and 33% of DEG were activated in *K. nitens* and *S. pratensis*, respectively (Fig. 5A). In contrast, the distribution of fold change amplitude values differed between the two charophytes where *S. pratensis* showed a general shift towards larger amplitudes of regulation (Fig. 5A). Even though the complement of auxin response proteins is different, all three bryophytes showed a similar pattern: 36-53% of DEG were activated, with very few genes showing an amplitude over 2-fold

up- or down-regulation (Fig. 5A). In contrast, 82% of DEG were activated in *C. richardii*. We also found that there was a notable difference in the distribution of fold-change values, with a larger fraction of genes being more strongly activated (maximum 28-fold; Fig. 5A).

We found that the number of auxin-responsive genes is positively correlated with the number of ARFs in land plants as seen in the expanded number of ARFs and DEG in *P. patens* and *C. richardii*. A switch to gene activation is not correlated with the number of ARFs, but rather with a duplication in the class A-ARFs in the ancestor of euphyllophytes and/or increase of Aux/IAA and TIR1/AFB. The increase in amplitude of auxin-dependent gene regulation in *C. richardii* could be a consequence of higher activation upon treatment, increased repression in the absence of auxin, or both. To determine its basis, we compared normalized expression values for the 20 top-most auxin-activated, and the 20 least auxin-activated genes in all species (Fig. 5B). This revealed that the increased amplitude of the top-most activated genes in *C. richardii* is not correlated with increased expression in the presence of auxin, but rather caused by reduced expression in its absence. This quantitative property of the auxin response system is correlated with the increased numbers of Aux/IAA genes.

Identification of a deeply conserved auxin-dependent gene set in land plants

Given that the mechanism of auxin response is ancient and conserved among all land plants, a key question is whether responses in different species involve regulation of a shared set of genes. To address this question, we performed tBLASTx searches among all DEG in our comparative transcriptome data and visualized the network of their similarities (Fig. S9 and Fig. S10). Even though BLAST filtering is not sufficient to distinguish orthology groups in large families such as kinases, we could identify several gene families to be commonly regulated by auxin in different land plants species. Classical primary auxin-responsive genes—the Aux/IAA, GH3 and SAUR families—were shown to be auxin responsive in many angiosperm species (Abel and Theologis, 1996). We found different bryophyte species to show auxin-dependence in only some of these three gene families (Fig. 6A), yet no species showed regulation of all three gene families. In contrast, *C. richardii* displayed auxin-dependence of members of all three gene families (Fig. 6A). Given that the Aux/IAA and GH3 proteins themselves regulate auxin levels or response, this indicates that a robust feedback mechanism involving all these gene families did not exist prior to the emergence of vascular plants, and bryophytes might have different feedback mechanism.

In addition, we identified the members of class II homeodomain-leucine zipper (C2HDZ) and WIP families to be commonly activated by auxin in all land plants in our RNA-Seq (note that no WIP gene was identified in the *A. agrestis* assembly). Indeed, qPCR analysis confirmed auxin-activation of C2HDZ (Fig. 6B). We also identified the members of auxin biosynthesis gene YUC family to be commonly down-regulated among multiple land plant species (except *A. agrestis*), and qPCR analysis demonstrated this to be true in *A. agrestis*, as well (Fig. 6B). It is known that some members of C2HDZ, WIP, and YUC families in *A. thaliana* are also up- or down-regulated by auxin respectively (Crawford et al., 2015; Sawa et al., 2002;



Takato et al., 2017). While homologs of C2HDZ were detected in the charophyte assemblies, none was regulated by auxin, which supports the different nature of the auxin response system in these species. In summary, land plants share a deeply conserved set of auxin up- and down-regulated genes.

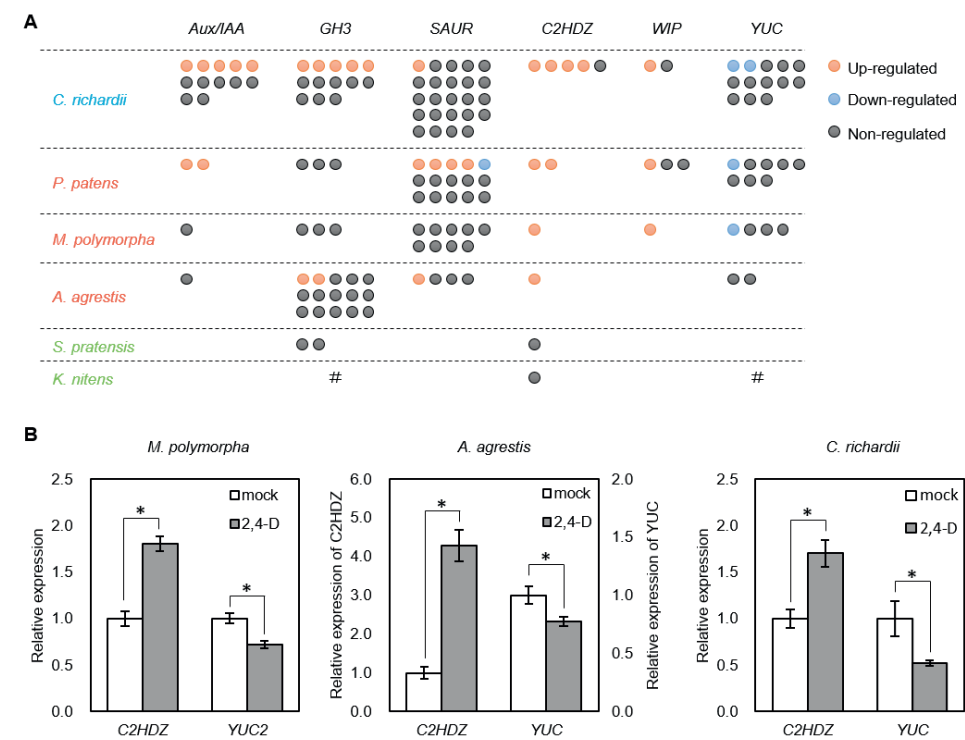


Figure 6: Identification of deeply conserved auxin-responsive genes. (A) Auxin-dependence of six well-known angiosperm auxin-responsive gene families (top) surveyed from de novo assembly-based transcriptomes in 6 species. Each circle indicates a gene copy of each gene family. Red, blue and grey circle indicate up-, down- and non-regulated genes in response to auxin. #: no homologs were identified in our transcriptome possibly due to low expression, or they might be lost during evolution. (B) qPCR analysis of conserved auxin-responsive genes. Auxin treatment was performed in the same condition with RNA-Seq experiment (10 μ M 2,4-D for 1h). Relative expression values are normalized by the expression of EF1 α in *Marchantia polymorpha* or the amount of total RNA in *Anthoceros agrestis* and *Ceratopteris richardii*. Each bar indicates average of expression with SD (biological replicates ≥ 3). *: $p < 0.01$ (t-test).

Contributions of ancient components to auxin response

Our phylogenomic analysis identified several components that are deeply conserved, yet whose contributions to auxin response are unknown: two deeply conserved non-canonical auxin signalling components lack important domains (ncIAA and ncARF), while class C-ARFs diverged from all other ARFs in charophytes prior to establishment of the NAP. To investigate the biological roles of these genes, we chose the liverwort *M. polymorpha*, the only genetically tractable model plant encoding ncIAA, ncARF and C-ARF genes. We first addressed ncIAA and ncARF function and performed CRISPR/Cas9-mediated mutagenesis (Sugano et al., 2014) and

obtained two different alleles for each gene which presumably cause a loss-of-function by frame shift mutation (*nciaa-6*, *nciaa-10*, *ncarf-2*, *ncarf-10*; Fig. 7A and Fig. S11A, B, E).

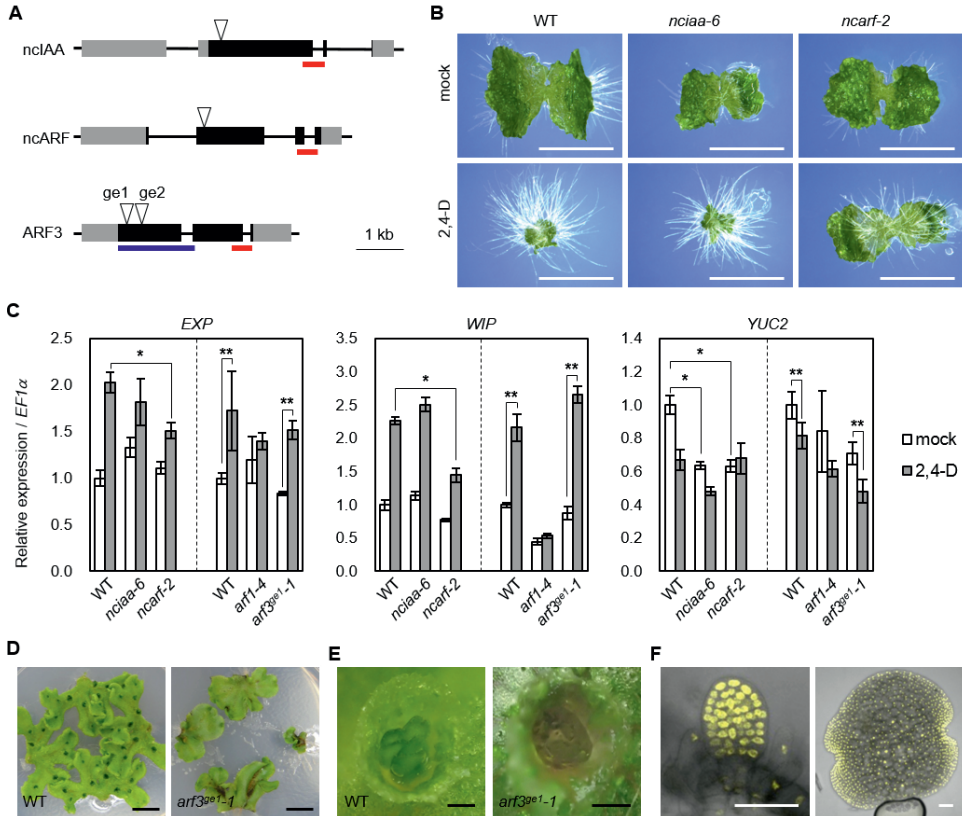


Figure 7: Genetic analysis of ancient components in *Marchantia polymorpha*. (A) Diagrams of gene structure and CRISPR/Cas9-mediated mutation in *nclAA*, *ncARF* and *ARF3* loci. Arrowheads indicate sgRNA target sites. Grey and black boxes indicate UTR and CDS, respectively. Red and blue bars indicate the region coding PB1 and DBD. (B) 10-day-old gemmalings grown without or with 3 μ M 2,4-D. Scale bars: 5 mm. (C) Expression analysis of auxin-responsive genes in WT, *nciaa*, *ncarf*, and *arf3* mutants by qPCR. 10-day-old gemmalings (*nciaa* and *ncarf*) or regenerating thalli (*arf1* and *arf3*) were treated with 10 μ M 2,4-D for 1 h. Each bar indicates average \pm SD (biological replicates = 3). Asterisks indicate significant differences: *; p < 0.01 (Tukey test), **; p < 0.05 (t-test). (D, E) Thallus tips grown for 2 weeks (D) and gemma cups (E) of WT and *arf3ge1-1* mutant. *arf3ge1-1* showed growth retardation and no mature gemmae, similar to the other alleles. (F) Expression analysis of proARF3:ARF3-Citrine in *arf3ge2-1* background. Left and right panel show developing and mature gemmae, respectively. Scale bars: 5 mm in (B and D), 0.5 mm in (E), 50 μ m in (F).

To investigate whether *nclAA* and *ncARF* are involved in auxin response, we grew mutants on auxin-containing medium. Exogenously supplied auxin causes severe inhibition of thallus growth and increased formation of rhizoids in wild-type (Fig. 7B; Ishizaki et al., 2012). *nciaa* mutants showed auxin response similar to wild-type, while growth inhibition was strongly suppressed in *ncarf* mutants although rhizoid formation was still promoted by auxin (Fig. 7B). We next selected two auxin-up-regulated genes (*EXP* and *WIP*) and one auxin-downregulated gene (*YUC2*; Eklund et al., 2015), and examined their expression in all mutants by qPCR analysis

(Fig. 7C). In *nciaa* mutants, the expression of auxin-upregulated genes responded similarly to the wild-type, while the expression of the auxin-repressed YUC2 gene was significantly reduced in the absence of auxin, but similarly repressed by auxin. In *ncarf* mutants, the basal expression of auxin-upregulated genes was similar to WT, while the expression after auxin treatment was significantly reduced in the mutants. The expression of YUC2 was reduced in mock condition and auxin treatment did not change the expression. Thus, in *M. polymorpha*, ncIAA may have a function in gene expression, but is not critical for auxin response itself. On the other hand, ncARF represents a novel positive regulator of both auxin-dependent gene activation and repression.

Finally, we focused on C-ARF function. While partial mutants have been reported in *A. thaliana*, no plants completely lacking C-ARF have been described. We used CRISPR/Cas9 gene editing to generate a series of loss-of-function mutants in MpARF3, the single C-ARF of *M. polymorpha* (*arf3ge1-1*, *arf3ge1-2*, *arf3ge2-1*; Fig. 7A and Fig. S11C, D). All three *arf3* mutants showed dramatic defects in development, notably in vegetative propagules (gemmae) which arrested before maturation, consistent with ubiquitous ARF3 protein accumulation in these structures (Fig. 7D-F and Fig. S11G). A previous study reported characterization of mutants in the class A-ARF in *M. polymorpha* (*arf1-4*) and showed that ARF1 is important for auxin response (Kato et al., 2017a). *arf1-4* produces narrower and twisted thallus which is distinct from flat thallus of *arf3* mutants. In addition, previous studies also showed that gemma development was regulated by Aux/IAA and the class A-ARF (Kato et al., 2015; Kato et al., 2017a), and we hence tested if transcriptional responses to auxin were altered in *arf3* mutants. Strikingly, all auxin-responsive genes we tested showed similar responses in WT and *arf3* mutants, while *arf1* mutants showed no auxin responses (Fig. 7C). This result suggests that, class C-ARF in *M. polymorpha* have different target genes from A-ARF and may not be critical for auxin-dependent gene regulation.

Discussion

Deep origin of nuclear auxin response in the ancestor of land plants

Phylogenetic analysis and domain structural analysis provided many insights into the origin of NAP and its evolutionary trajectory. All subdomains of dedicated auxin-response proteins were recovered in transcriptomes from red algae and chlorophytes, but the multidomain protein appears only in the charophyte and land plant lineage. These findings show that proto-ARF transcription factor was established during the evolution of ancestral charophytes by combining existing domains. However, given that no defined Aux/IAA and TIR1/AFB could be identified in charophytes, a complete nuclear auxin response system is limited to land plants. Ancestors of TIR1/AFB and COI1 co-receptors could be identified in charophytes, but detailed residue analysis suggested these to be neither auxin nor JA receptor. Thus, duplication of this gene, as well as multiple mutations in the LRR domain, must have preceded the deployment of these proteins as co-receptors. Auxin-dependence of ARFs is mediated by auxin-triggered degradation of Aux/IAA proteins, bridging ARF and TIR1/AFB proteins through two protein domains: the ARF-interacting PB1 domain and the TIR1/AFB-interacting domain II. We did find charophyte

PB1-containing proteins that form a sister clade of land plant Aux/IAA. However, domain II was not detected in these proteins. Along with innovations in the proto-TIR1/AFB/COI1 protein, gain of a minimal degron motif in the Aux/IAA precursor likely completed the auxin response system in the early ancestor of land plants. Whether proto-TIR1/AFB/COI1 interact with Aux/IAA-like protein via an unknown ligand would be an interesting question for future analysis.

Auxin responses in algal species

Despite the lack of defined Aux/IAA and TIR1/AFB auxin co-receptor, the charophytes *K. nitens* and *S. pratensis* showed an extensive transcriptional response to exogenously supplied 2,4-D within 1 hour. A recent independent study showed IAA-dependent gene expression in *K. nitens* upon prolonged treatment with higher concentrations (100 μ M for 10 h to 7 days; Ohtaka et al., 2017). While *S. pratensis* has a proto-C-ARF, *K. nitens* does not appear to have proto-ARFs. Thus, by definition this response system must be different from the land plant auxin response system. Indeed, the charophyte orthologue of core land plant auxin responsive genes (C2HDZ) did not respond to 2,4-D and IAA. There was little, if any, overlap between auxin-responsive transcripts in the two charophytes, and in qPCR experiments on individual genes we noticed a high variability between experiments (not shown). Thus, it appears that charophytes do respond to auxin-like molecules, but this response may not be robust, or it may strongly depend on growth conditions. Auxin resembles indole and tryptophan, and it is possible that the response to auxin observed is in fact a metabolic response to nutrient availability. Presence of endogenous IAA is observed in a wide range of algal species including charophytes, chlorophytes, rhodophytes, chromista, and cyanobacteria (Žižková et al., 2017). Moreover, non-photosynthetic bacteria and fungi produce IAA and use it for communication with plants and algae (Amin et al., 2015; Fu et al., 2015), and thus it is likely that a response mechanism independent of the NAP exists in these species.

Function of the ancestral ARFs

Our data clearly indicate that ARF transcription factors were established in common ancestor of charophytes and land plants. Structural homology models suggest that all the important residues for DNA-binding are conserved in proto-ARFs and these may bind the same target DNA sequences. This should be assessed by biochemical experiments in the future. Given that there is a core set of auxin-regulated genes shared in all land plants, an intriguing possibility is that proto-ARFs already regulated this core set of genes that only became auxin-dependent upon establishment of TIR1/AFB and Aux/IAA proteins. Identification of the transcriptional targets of these proto-ARFs should help address this question. In any event, proto-ARFs—as well as critical residues for DNA binding—have been retained in many algal genomes for hundreds of millions of years, which suggests that they perform a biologically relevant function. Whether this function is related to the processes that auxin controls in land plants is an open question.

Interestingly, our phylogenetic analysis indicated that the split between class C- and A/B-

ARFs occurred in charophytes before the establishment of Aux/IAA-TIR1/AFB co-receptor, and by extension likely before proto-ARFs were auxin-dependent. This suggests that class C-ARFs are fundamentally different from class A/B-ARFs. Indeed, genetic analysis in *M. polymorpha* revealed that its C-ARF likely does not act in auxin-dependent gene regulation. Several studies in *A. thaliana* showed that C-ARFs are involved in auxin response but the proposed role was different between studies (Ding and Friml, 2010; Liu et al., 2010; Mallory et al., 2005; Wang et al., 2005; Yang et al., 2013). In addition, C-ARFs of *A. thaliana* generally have weak affinity to Aux/IAA proteins (Piya et al., 2014). To clarify the function of this ancient ARF subfamily, auxin-responsiveness of C-ARF proteins and relationship with A- or B-ARFs should be investigated in different species.

Novel components in auxin response

A surprising outcome of the phylogenomic analysis was the discovery of two deeply conserved non-canonical proteins: ncIAA and ncARF. Charophytes have an Aux/IAA-like protein containing a PB1 domain, but lacking domain II, which is critical for auxin perception. This protein could regulate the function of proto-ARF (or proto-RAV), but not in an auxin-dependent manner. While the canonical Aux/IAA gave rise to a large gene family, the ncIAA clade represented by a single member in every evolutionary node. The retention of a single ncIAA gene across plants suggests a fundamental function. Unfortunately, our mutant analysis in *M. polymorpha* could not reveal the function of ncIAA in auxin response and development in vegetative phase. ncIAA might have a function only in other developmental stages, or under specific stress conditions or environmental signals. No mutant in the *Arabidopsis* IAA33 gene has yet been reported, and perhaps such a mutant will help understand the ancient function of this protein.

This work revealed that a class A-ARF-derived ncARF subfamily lacking a DBD is evolutionarily conserved among liverworts, mosses, and lycophytes. Mutant analysis using *M. polymorpha* clearly showed that ncARF functions as positive regulator in transcriptional auxin responses. There are two hypothetical models for ncARF function. 1) ncARF protects canonical ARFs from AUX/IAA-mediated inactivation through the interaction of PB1 domain. 2) ncARF interacts with target gene loci by interaction with canonical ARFs and help activate expression by recruiting co-factors. Irrespective of the mechanism of ncIAA and ncARF function, future models of auxin response will need to incorporate these conserved components.

Functional impact of increased complexity in NAP components

Through comparative transcriptomics we infer that the number of DNA-binding ARF transcription factors scales with the number of auxin-regulated genes. Both *P. patens* and *C. richardii* have an expanded set of ARFs and display substantially more auxin-responsive genes than *A. agrestis* and *M. polymorpha*. It is likely that later duplications in the ARF family in the seed plants led to the thousands of auxin-responsive genes in these species (Paponov et al., 2008). Another key evolutionary change is the transition from mostly gene repression to gene

activation. We infer that this transition occurred in a common ancestor of euphyllophytes, and transcriptome analysis in *A. thaliana* and *O. sativa* shows this pattern persists in angiosperms (Jain and Khurana, 2009; Paponov et al., 2008). There is a defining difference between bryophyte and euphyllophyte ARF families—a persisting duplication in the class A-ARFs. We hypothesize that the euphyllophytes duplication created an ARF copy that is more potent, or perhaps even specialized for gene activation. However, we cannot exclude the possibility that the difference in endogenous auxin levels or tissue complexity among species may result in different sensitivity to auxin treatment.

The comparative transcriptomics also adds an interesting twist to our understanding of the functional distinction among ARF classes. Class A-ARFs are considered activators, and class B-ARFs repressors, perhaps through competing with class A-ARFs (Lavy et al., 2016; Ulmasov et al., 1999). Despite a complete lack of class B-ARFs, the hornwort *A. agrestis* showed comparable auxin-dependent gene repression to the other bryophytes, suggesting that auxin-dependent gene repression may not be mediated by class B-ARFs. Based on these findings, the role of class B-ARFs in auxin response may need to be reconsidered.

A remarkable difference between bryophyte and euphyllophyte auxin-dependent transcriptomes is the appearance of genes with a large amplitude of regulation in the latter. Many auxin-responsive genes that were first identified in angiosperms such as *A. thaliana* have very high amplitudes (Lee et al., 2009), but this appears to be a later innovation in the response system. The high amplitude is caused by more effective repression of gene activity in the no-auxin state, a property that is likely mediated by Aux/IAA proteins. Indeed, ferns have a much larger set of Aux/IAA proteins, as do all seed plants, and we propose that expansion of the Aux/IAA family enabled plants to articulate a clear distinction between on and off states in auxin response. In summary, this analysis reveals several design principles of the auxin response system.

Materials and Methods

Plant materials and culture condition

Male *M. polymorpha* strain Takaragaike-1 (Tak-1) was used as wild type and cultured as described previously (Kato et al., 2015). *K. nitens* (NIES-2285), *P. patens* (Gransden), and *A. agrestis* (Oxford) were cultured on BCD medium (Cove et al., 2009) solidified with 1% agar under the same condition with *M. polymorpha*. *S. pratensis* (UTEX928) was cultured on Guillard's Woods Hole medium (Nichols, 1973), pH7.9 containing 1% agar under white light with a 16-h light/8-h dark cycle at 22°C. *C. richardii* (Hn-n) was cultured on C-fern medium (Plackett et al., 2015) under continuous white light at 28°C.

Data used

Data access to 1000 plant transcriptomes was provided by the OneKP consortium (www.onekp.com; Matasci et al., 2014). All the transcriptome assemblies of the species from red algae, green algae, bryophytes, lycophytes, monilophytes, gymnosperms and basal angiosperms that were safely



identified as non-contaminated has been used for this analysis (Supplementary file 1). CDS and protein sequences encoding all the orthologous genes in the three (ARF, Aux/IAA and TIR1/AFB) gene families from *M. polymorpha*, *P. patens*, *Amborella trichopoda*, *Oryza sativa*, *Zea mays*, *Solanum lycopersicum* and *A. thaliana* were obtained from Phytozome ver11 (phytozome.jgi.doe.gov/pz/portal.html). Aux/IAA genes from *Picea abies* were obtained from Spruce Genome Project (www.congenie.org). *K. nitens* genome information was accessed from *Klebsormidium nitens* NIES-2285 genome project (Hori et al., 2014).

Phylogeny construction

After extracting all the homologous sequences using the method mentioned earlier (Chapter 2), all the sequences were further tested by BLASTx search against *A. thaliana* proteome to confirm orthology inferences. Some PB1 domain sequences in chlorophytes that showed low similarity to *A. thaliana* proteins were also compared with *M. polymorpha* sequences to ascertain orthology. MAFFT iterative refinement algorithm (E-INS-i) was used to align the CDS sequences. Alignment positions with more than 50% gaps were removed using the Phyutility program (ver2.2.6; <http://blackrim.org/programs/phyutility/>) before the phylogeny construction. PartitionFinder (ver1.1.1; Lanfear et al., 2012) was used to identify the most suitable evolutionary model for all the three gene families using the complete trimmed alignments on all the domains. Maximum likelihood algorithm implemented in RAxML (ver8.1.20; Stamatakis, 2014) with General Time Reversible (GTR) model of evolution under GAMMA rate distribution with bootstopping criterion (up to a maximum of 1000 bootstraps) was used for the phylogenetic analysis. Obtained trees were visualized using the iTOL (<http://itol.embl.de/>) phylogeny visualization program. Phylogenetic trees were cleaned up manually for misplaced sequences as well as for clades with long branch attraction.

Auxin treatment

M. polymorpha gemmae or thallus explant without meristem and *A. agrestis* small thalli were planted on the medium covered with nylon mesh (100 µm pore) and grown for 10 days. *P. patens* protonematal tissues were grown on the medium covered with cellophane for 10 days. Sterilised spores of *C. richardii* were grown for 2 weeks after which fertilization was performed by adding 5 ml of water on the plate. 7 days after fertilization, prothalli carrying sporophytic leaves were transferred on the medium covered with nylon mesh and grown for a further 7 days, after which sporophytes contained 3–4 leaves. After growing, plants with mesh or cellophane were submerged into liquid medium and cultured for 1 day. After pre-cultivation, 2,4-D was added to a final concentration of 10 µM and plants were incubated for 1h. Excess liquid medium were removed with paper towels and plants were frozen in liquid nitrogen. *K. nitens* and *S. pratensis* were streaked on solid medium and grown for 2 weeks. Algal cells were collected into 40 ml of liquid medium and cultured for 1 day with shaking at ~120 rpm. Then 2,4-D was added so that final concentration became 10 µM, followed by incubation for 1 h with shaking. After auxin treatment, algal cells were collected using filter paper and frozen in liquid nitrogen.

RNA extraction and sequencing

Frozen plant sample were grinded into fine powder with mortar and pestle. RNA from *K. nitens*, *S. pratensis*, *M. polymorpha*, and *P. patens* were extracted using Trizol Reagent (Thermo Fisher Scientific) and RNeasy Plant Mini Kit (QIAGEN). RNA from *A. agrestis* and *C. richardii* were extracted using Spectrum™ Plant Total RNA Kit (Sigma-Aldrich). Total RNA was treated with RNase-free DNase I set (QIAGEN) and purified with RNeasy MinElute Clean Up Kit (QIAGEN). RNA-Seq library construction with TruSeq kit (Illumina) and 100 bp paired-end sequencing with Hiseq4000 (Illumina) were performed by BGI TECH SOLUTIONS (HONGKONG).

Quantitative RT-PCR

cDNA was synthesized with iScript cDNA Synthesis Kit (Bio-Rad). Quantitative PCR was performed using iQ™ SYBR® Green Supermix (Bio-Rad) and CFX384 Touch™ Real-Time PCR Detection System. A two-step cycle consisting of denaturation at 95°C for 10 seconds followed by hybridization/elongation at 60°C for 30 seconds, was repeated 40 times and then followed by a dissociation step. Three technical and biological replicates were performed for each condition. PCR efficiencies were calculated using CFX Manager (Bio-Rad) software in accordance with the manufacturer's instructions. For *M. polymorpha*, relative expression values were normalized by the expression of EF1 α (Saint-Marcoux et al., 2015). All primers used for the analysis are listed in Supplementary file 3.

RNA-Seq data analysis

Obtained raw fastq reads were checked for quality control using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc). De novo transcriptome assemblies for all 6 species were generated using Trinity (<http://trinityrnaseq.github.io>) with default settings. To avoid any possible contamination from sequencing method and to improve the data quality, raw reads from land plants were mapped against charophyte de novo assemblies using Bowtie2 (<http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>) in default settings and all the perfectly mapped pairs were removed, after which new assemblies were generated from pure raw read data for each species. In a similar way, contamination was removed in charophytes by mapping them against land plant transcriptome assemblies. Once the pure de novo transcriptome assemblies were generated, again Bowtie2 was used to map individual sample to the respective transcriptome assemblies using default parameters. Further, to improve the read count estimation and reduce the redundancy in Trinity transcripts, Corset (Davidson and Oshlack, 2014) was implemented to estimate raw read counts using the Bowtie2 mapped alignment data. The obtained raw read counts were normalized and differentially expressed genes (Padj < 0.01) were identified using DEseq2 (Love et al., 2014) implemented in R Bioconductor package. All the RNAseq raw reads were deposited in NCBI Short Read Archive (SRA) under the BioProjectID: PRJNA397394 (www.ncbi.nlm.nih.gov/bioproject/397394).

Alignments and homology modelling

All other protein alignments mentioned in the manuscript were generated using ClustalOmega (<http://www.ebi.ac.uk/Tools/msa/clustalo/>). Visualization of the alignments were generated using Esript (esript.ibcp.fr). Homology models were generated using Modeller v9.17 (<https://salilab.org/modeller/>). Modelled 3D structures were visualized using PyMol v1.7.4 (The PyMOL Molecular Graphics System, Schrödinger, LLC).

Core auxin-responsive gene set

All the up-regulated genes' nucleotide sequences from the 6 species were aligned against the same sequences using tBLASTx to find the similar (orthologous) genes among various species. From these results, the BLAST hits with E-value less than 0.001 with a length of at least 30 amino acids were considered for further analysis. Moreover, these sequences were also searched for orthologues in *A. thaliana* proteome using BLASTx. Both the similarities among the six species and the orthologous gene information from *A. thaliana* were loaded into Cytoscape (www.cytoscape.org) to visualize the network of similar gene families. A similar procedure was performed for finding the commonly downregulated gene families.

Mutant generation for *M. polymorpha*

To generate the entry clones carrying sgRNA cassette, pairs of oligo DNAs (HK001/HK002 or HK003/HK004 for ARF3, HK162/HK163 for ncARF, HK168/HK169 for ncIAA) were annealed and cloned into pMpGE_En03 (Addgene) using BsaI site. The sequence of oligo DNAs are listed in Supplementary file 3. Resultant sgRNA cassette were transferred into pMpGE_010 (Addgene) by LR reaction using Gateway® LR Clonase® II Enzyme Mix (Thermo Fisher Scientific). Transformation into Tak-1 was performed as described previously (Kubota et al., 2013) using Agrobacterium strain GV3101:pMp90. For genotyping, genomic DNA was extracted by simplified CTAB (cetyltrimethylammonium bromide) method (<http://moss.nibb.ac.jp/protocol.html>). Genomic region including target site of sgRNA was amplified with PCR using the primer set HK079/HK131 (ARF3), HK172/HK173 (ncARF) and HK174/HK175 (ncIAA), and sequenced. All primers used in this study are listed in Supplementary file 3.

Expression analysis of MpARF3 protein

MpARF3 promoter fragment including 5' UTR and 3 kb up stream region was amplified with PCR using the primer set HK111/HK026 and cloned into pMpGWB307 (Ishizaki et al., 2015) using XbaI site (pJL002). Genomic CDS of MpARF3 without stop codon was amplified with PCR using the primer set HK027/028 and subcloned into pENTR/D-TOPO vector (Thermo Fisher Scientific). Mutation which confers resistant to sgRNA was introduced by PCR using primer set HK137/138. Then mutated CDS fragment was transferred into pJL002 by LR reaction and fused with promoter and Citrine tag (pHKDW103). All primers used in this study

are listed in Supplementary file 3. Resultant vector was transformed into *arf3ge2-1* mutant thallus as described previously. Citrine signal and bright field images were captured using a Leica SP5-II confocal laser scanning microscope system, with excitation at 514 nm and detection at 520–600 nm.

Supplementary files

All the supplementary files mentioned in this chapter are available online as “Additional files” under the link: <https://doi.org/10.7554/eLife.33399.022>.

Acknowledgments

We thank Dr. Jane A. Langdale for distributing plant materials of *A. agrestis* and *C. richardii*, and Jasper Lamers and Lisa Olijslager for contributing to MpARF3 analysis. We are grateful to all contributors of the OneKP project for generating a comprehensive transcriptome database, and Eric Carpenter for providing access. We thank Dr. Kuan-Ju Lu and Dr. Nicole van ‘t Wout Hofland for helpful comments on the manuscript. This study was supported by an EMBO Long-Term Postdoctoral Fellowship (ALTF 415-2016) to H.K. and a VICI grant from the Netherlands Organization for Scientific Research (NWO; 865.14.001) to D.W.

Supplementary information

Supplementary file 1: Species used in phylogenomic analysis.

Supplementary file 2: Multiple sequence alignments used in the study.

Supplementary file 3: Primers used in this study.

Supplementary file 4: Detail for network of Up-regulated genes.

Supplementary file 5: Detail for network of Down-regulated genes.



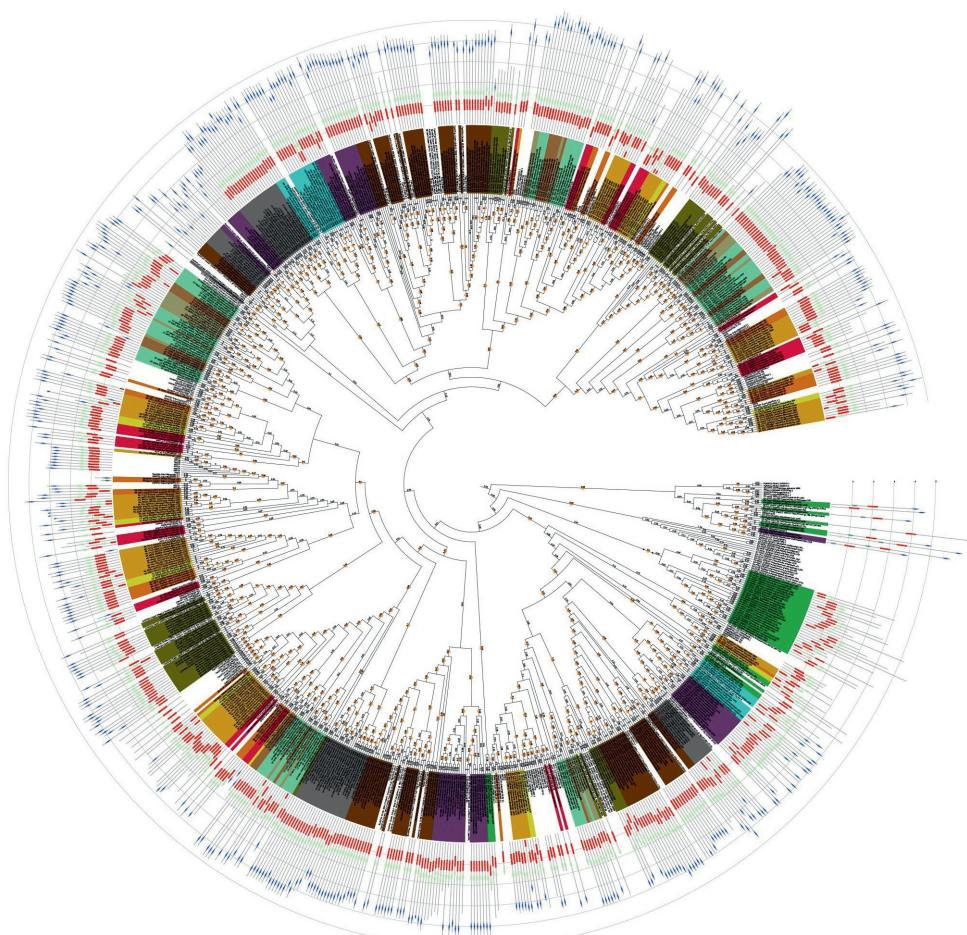


Figure S1: Phylogenetic tree of ARF and RAV proteins. Label color shows the taxonomic group of each protein as indicated in Figure S4. Numbers along with the branches indicate branch length. Orange circles indicate the bootstraps higher than 75. Colored boxes connected with gray bar shows the domain structure of each protein. Red: B3, green: DD2+AD, blue: PB1, gray: AP2. The complete tree can be found at interactive Tree of Life (iTOL): <http://itol.embl.de/shared/dolfweijers>

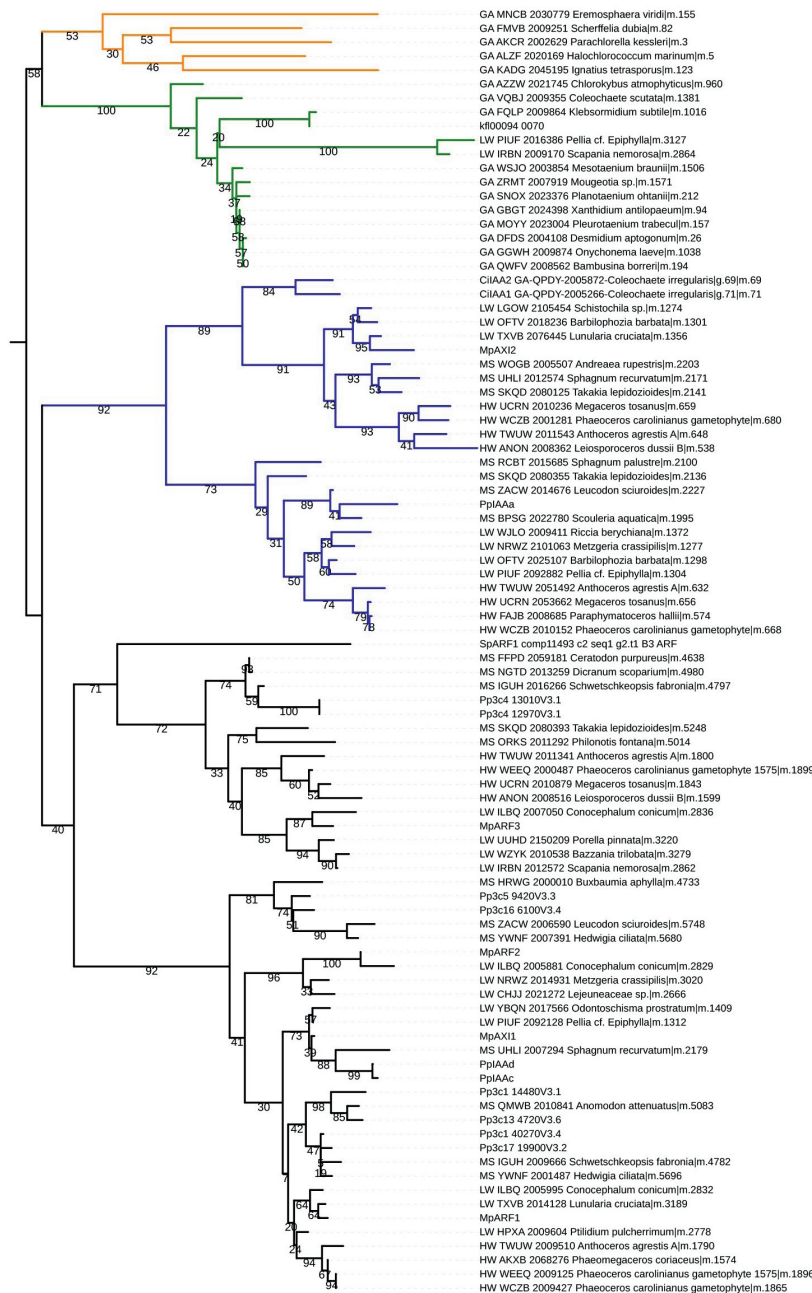


Figure S3: Phylogenetic tree based on PB1 domain. Colored branches indicate protein families. Orange: Chlorophytes, green: proto-RAV, blue: Aux/IAA, black: (proto)-ARF. Numbers along with the branches indicate bootstrap values. The complete tree can be found at interactive Tree of Life (iTOL): <http://itol.embl.de/shared/dolfweijers>

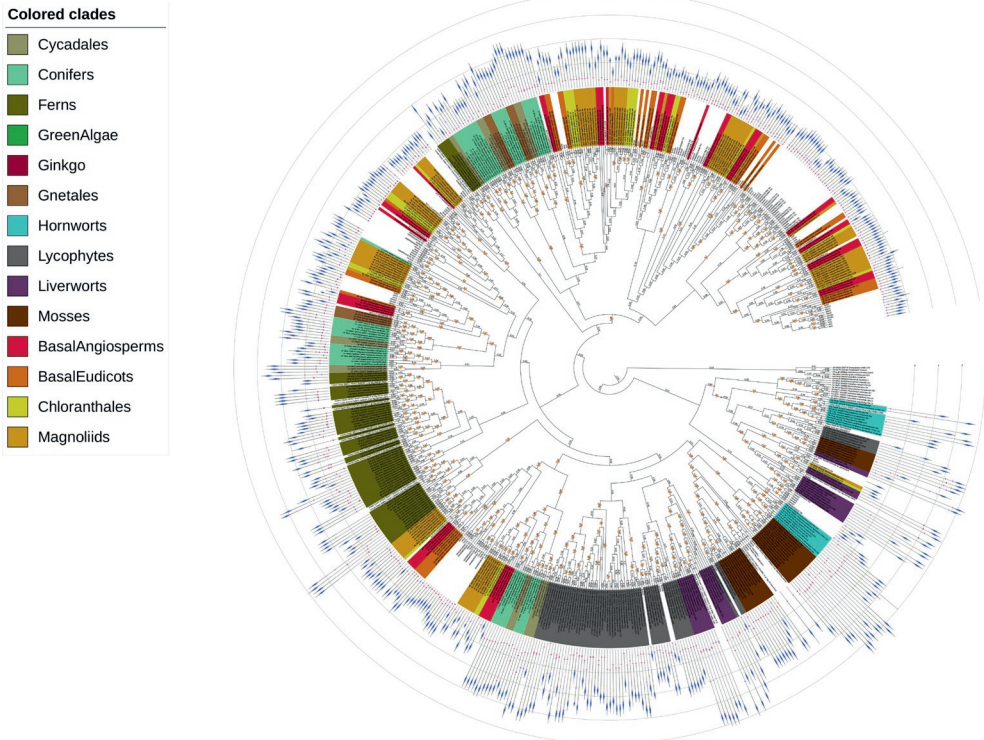


Figure S4: Phylogenetic tree of Aux/IAA. Label color shows the taxonomic group of each protein as indicated in top. Colored boxes connected with gray bar shows the domain structure of each protein. Magenta: domain I, yellow domain II, blue: PB1. Numbers along with the branches indicate branch length. Orange circles indicate bootstrap values higher than 75. The complete tree can be found at interactive Tree of Life (iTOL): <http://itol.embl.de/shared/dolfweijers>.

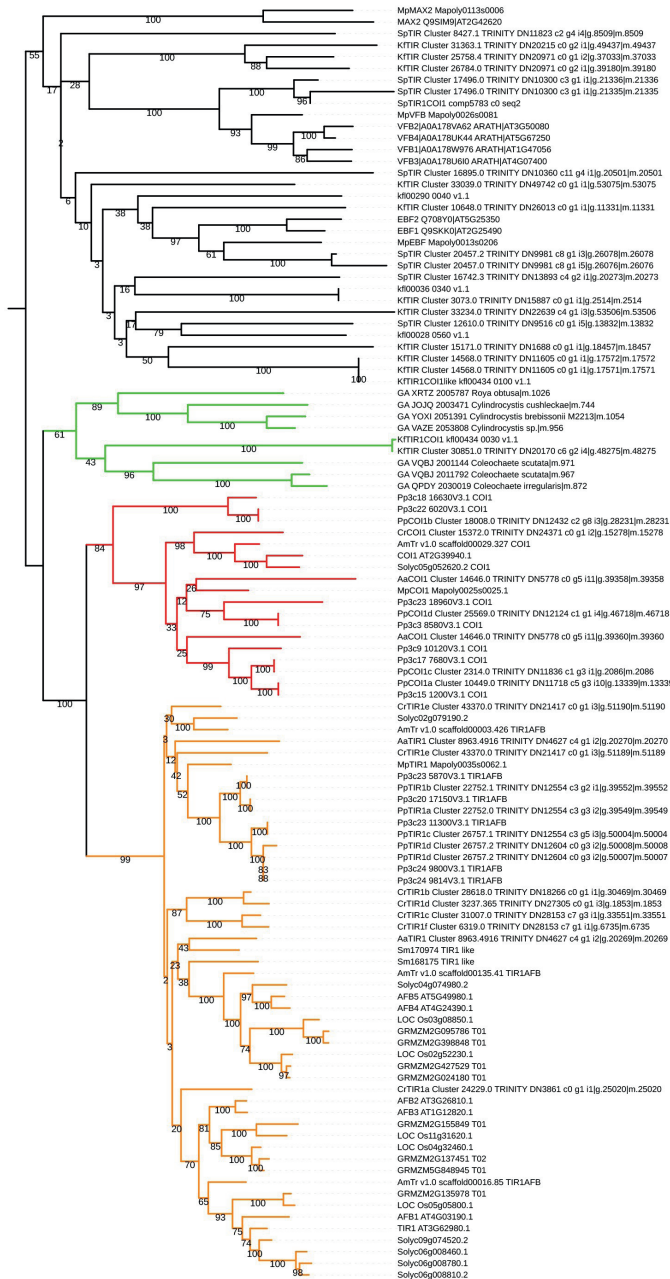


Figure S5: Phylogenetic tree of the proteins containing F-box and LRR. Colored branches indicate protein families. Green: TIR1/COI1 precursor of Charophytes, red: COI1, orange: TIR1/AFB, black: the others. Numbers along with the branches indicate bootstrap values. The complete tree can be found at interactive Tree of Life (iTOL): <http://itol.embl.de/shared/dolfweijers>

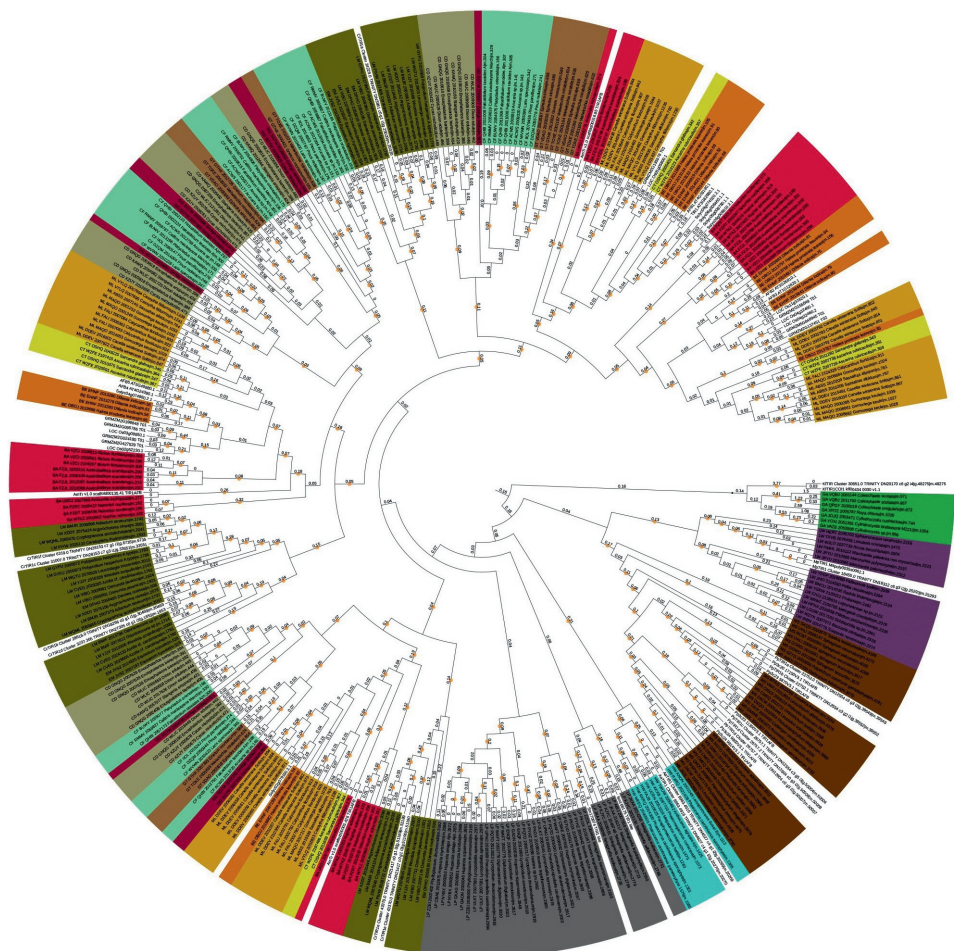


Figure S6: Phylogenetic tree of TIR1/AFB. Label color shows the taxonomic group of each protein as indicated in Figure S4. Numbers along the branches indicate branch length. Orange circles indicate bootstrap values higher than 75.

	De novo assembly-based		Genome-based		% Match (b/a)
	Number of DEG	Matches to genome data (a)	Number of DEG	Matches to de novo data (b)	
<i>K. nitens</i>	1094	970	1265	874	90.1
<i>M. polymorpha</i>	105	98	98	67	68.4
<i>P. patens</i>	1090	1035	1138	840	81.2

Figure S7: Number of DEG in de novo assembly- or genome-based transcriptome analysis.

Species	Trinity Assembly Length (bp)	Corset Assembly Length (bp)	Trinity Transcripts	Corset Transcripts	Trinity Genes	Corset Genes (Clusters)	GC %	DEG (up)	DEG (down)	DEG (total)
<i>K. nitens</i>	103,789,047	96,017,668	90,079	62,718	62,978	36,209	55	409	685	1,094
<i>S. pratensis</i>	73,580,548	67,121,353	79,576	55,877	47,425	26,062	39	563	1,118	1,681
<i>A. agrestis</i>	305,335,425	302,084,291	155,454	143,697	43,514	36,512	51	65	94	159
<i>M. polymorpha</i>	127,811,269	121,162,080	87,527	63,666	52,502	30,749	48	56	49	105
<i>P. patens</i>	294,940,461	288,755,562	163,952	141,424	52,712	36,521	46	395	695	1,090
<i>C. richardii</i>	208,082,319	190,903,669	191,760	130,844	108,289	57,950	43	339	74	413

Figure S8: Summary statistics of comparative RNA-Seq analysis.

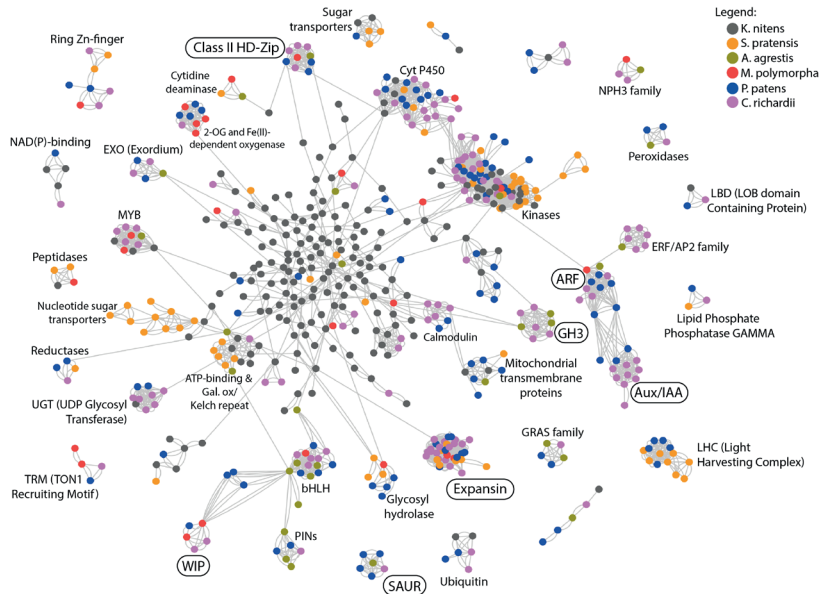


Figure S9: Network of Up-regulated genes shared between different species upon auxin treatment. Nodes represent the genes and edges represent the presence of BLAST similarity. Colors indicate the species in the legend above. Note that two edges connect nodes if the genes are bi-directional BLAST hits. See also Supplementary file 4.

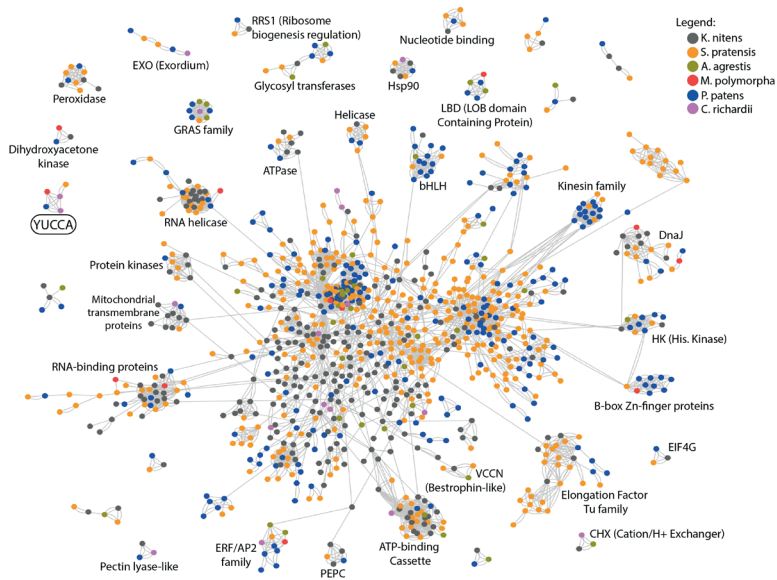


Figure S10: Network of Down-regulated genes shared between different species upon auxin treatment. Nodes represent the genes and edges represent the presence of BLAST similarity. Note that two edges connect nodes if the genes are bi-directional BLAST hits. Colors indicate the species in the legend above. See also Supplementary file 5.

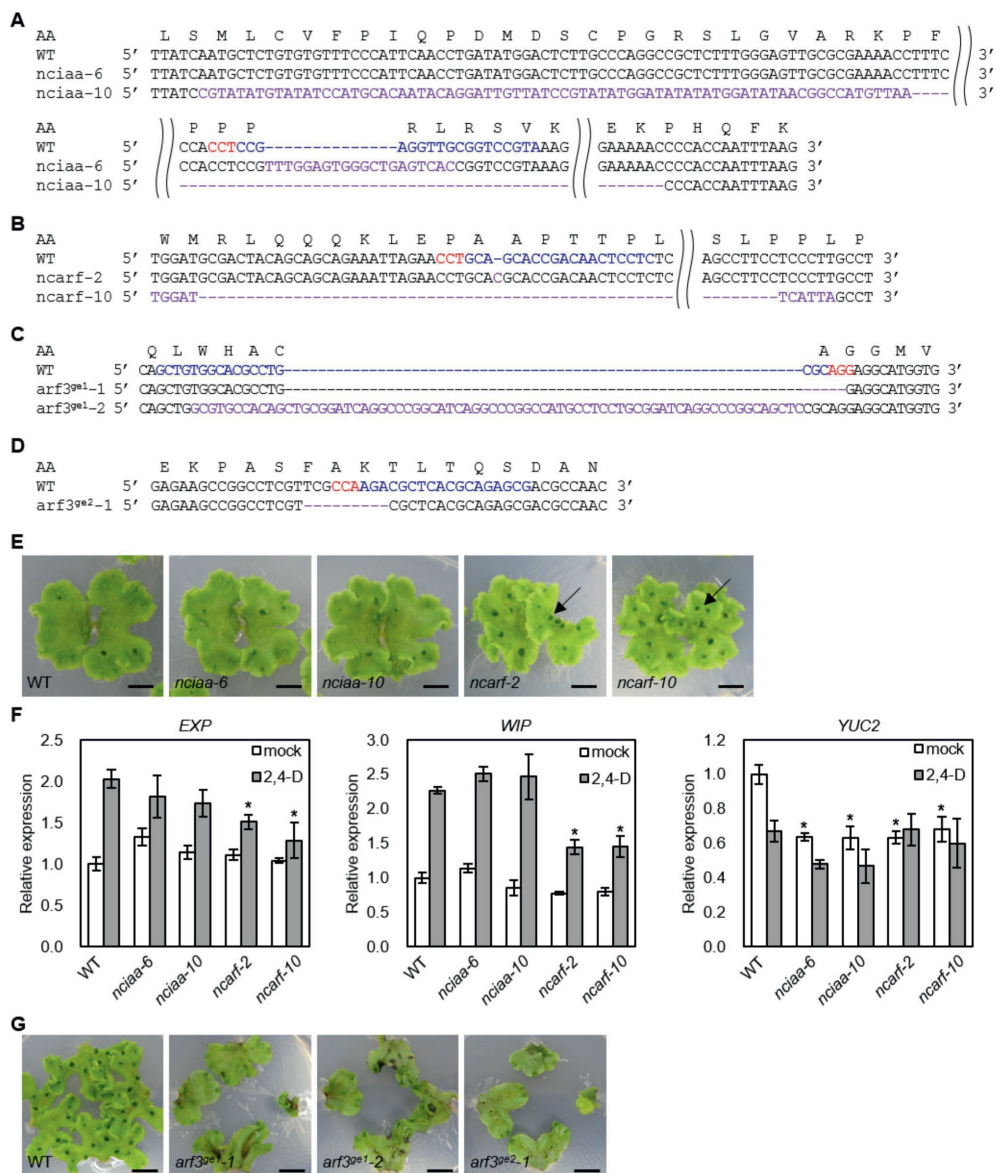


Figure S11: CRISPR/Cas9-mediated mutagenesis in *M. polymorpha*. (A-D) Mutations detected by sequencing analysis. The amino acid (AA) sequences encoded in WT are shown at the top. WT sequence is shown with the PAM sequence highlighted in red and the target sequence of sgRNA in blue. Purple bases indicate mutation. *nciaa-6*: 6 bp deletion and 20 bp insertion, *nciaa-10*: 776 bp deletion and 75 bp insertion, *ncarf-2*: 1 bp insertion, *ncarf-10*: 486 bp deletion and 6bp insertion, *arf3^{ge1-1}*: 5 bp deletion, *arf3^{ge1-2}*: 11 bp deletion and 72 bp insertion, *arf3^{ge2-1}*: 9 bp deletion. (E) 3-week-old gemmaling. Arrows indicate the thalli formed with up-side-down. (F) qPCR analysis on 10-day-old gemmalings with or without 10 μ M 2,4-D treatment for 1 h. Relative expression values are normalized by the expression of EF1 α . Each bar indicates average with SD (biological replicate = 3). Each asterisk indicates significant difference between WT and mutants in the same condition ($p < 0.01$, Tukey test). (G) Thallus tips of WT and *arf3* mutants grown for 2 weeks. Scale bars = 5 mm.

References

- Abel, S., and Theologis, A. (1996). Early genes and auxin action. *Plant Physiol.* 111, 9-17.
- Amin, S.A., Hmelo, L.R., van Tol, H.M., Durham, B.P., Carlson, L.T., Heal, K.R., Morales, R.L., Berthiaume, C.T., Parker, M.S., Djunaedi, B., et al. (2015). Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* 522, 98-101.
- Bates, G.W., and Goldsmith, M.H. (1983). Rapid response of the plasma-membrane potential in oat coleoptiles to auxin and other weak acids. *Planta* 159, 231-237.
- Boer, D.R., Freire-Rios, A., van den Berg, W.A., Saaki, T., Manfield, I.W., Kepinski, S., López-Vidrio, I., Franco-Zorrilla, J.M., de Vries, S.C., Solano, R., et al. (2014). Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. *Cell* 156, 577-589.
- Cove, D.J., Perroud, P.F., Charron, A.J., McDaniel, S.F., Khandelwal, A., and Quatrano, R.S. (2009). Culturing the moss *Physcomitrella patens*. *Cold Spring Harb. Protoc.* 2009, pdb prot5136.
- Crawford, B.C., Sewell, J., Golembeski, G., Roshan, C., Long, J.A., and Yanofsky, M.F. (2015). Genetic control of distal stem cell fate within root and embryonic meristems. *Science* 347, 655-659.
- Davidson, N.M., and Oshlack, A. (2014). Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol.* 15, 410.
- Dharmasiri, N., Dharmasiri, S., and Estelle, M. (2005). The F-box protein TIR1 is an auxin receptor. *Nature* 435, 441-445.
- Ding, Z., and Friml, J. (2010). Auxin regulates distal stem cell differentiation in Arabidopsis roots. *Proc. Natl. Acad. Sci. USA* 107, 12046-12051.
- Eklund, D.M., Ishizaki, K., Flores-Sandoval, E., Kikuchi, S., Takebayashi, Y., Tsukamoto, S., Hirakawa, Y., Nonomura, M., Kato, H., Kouno, M., et al. (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.
- Etherton, B. (1970). Effect of Indole-3-acetic Acid on Membrane Potentials of Oat Coleoptile Cells. *Plant Physiol.* 45, 527-528.
- Finet, C., Berne-Dedieu, A., Scutt, C.P., and Marlétaz, F. (2013). Evolution of the ARF gene family in land plants: old domains, new tricks. *Mol. Biol. Evol.* 30, 45-56.
- Flores-Sandoval, E., Eklund, D.M., and Bowman, J.L. (2015). A Simple Auxin Transcriptional Response System Regulates Multiple Morphogenetic Processes in the Liverwort *Marchantia polymorpha*. *PLoS Genet.* 11, e1005207.
- Fu, S.F., Wei, J.Y., Chen, H.W., Liu, Y.Y., Lu, H.Y., and Chou, J.Y. (2015). Indole-3-acetic acid: A widespread physiological code in interactions of fungi with other organisms. *Plant Signal. Behav.* 10, e1048052.
- Gray, W.G., Kepinski, S., Rouse, D., Leyser, O., and Estelle, M. (2001). Auxin regulates SCFTIR1-dependent degradation of AUX/IAA proteins. *Nature* 414, 271-276.
- Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., et al. (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* 5, 3978.
- Ishizaki, K., Nishihama, R., Ueda, M., Inoue, K., Ishida, S., Nishimura, Y., Shikanai, T., and Kohchi, T. (2015). Development of Gateway Binary Vector Series with Four Different Selection Markers for the Liverwort *Marchantia polymorpha*. *PLoS one* 10, e0138876.
- Ishizaki, K., Nonomura, M., Kato, H., Yamato, K.T., and Kohchi, T. (2012). Visualization of auxin-mediated transcriptional activation using a common auxin-responsive reporter system in the liverwort *Marchantia polymorpha*. *J. Plant Res.* 125, 643-651.
- Jain, M., and Khurana, J.P. (2009). Transcript profiling reveals diverse roles of auxin-responsive genes during reproductive development and abiotic stress in rice. *FEBS J.* 276, 3148-3162.
- Kato, H., Ishizaki, K., Kouno, M., Shirakawa, M., Bowman, J.L., Nishihama, R., and 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 Genet.* 11, e1005084.
- Kato, H., Kouno, M., Takeda, M., Suzuki, H., Ishizaki, K., Nishihama, R., and Kohchi, T. (2017a). The Roles of the Sole Activator-Type Auxin Response Factor in Pattern Formation of *Marchantia polymorpha*. *Plant Cell Physiol.* 58, 1642-1651.
- Kato, H., Nishihama, R., Weijers, D., and Kohchi, T. (2017b). Evolution of nuclear auxin signaling: lessons from genetic studies with basal land plants. *J. Exp. Bot.* doi: 10.1093/jxb/erx267.
- Katsir, L., Schilmiller, A.L., Staswick, P.E., He, S.Y., and Howe, G.A. (2008). COI1 is a critical component of a receptor for jasmonate and the bacterial virulence factor coronatine. *Proc. Natl. Acad. Sci. US A* 105, 7100-7105.
- Kepinski, S., and Leyser, O. (2005). The Arabidopsis F-box protein TIR1 is an auxin receptor. *Nature* 435, 446-451.
- Kim, J., Harter, K., and Theologis, A. (1997). Protein-protein interactions among the Aux/IAA proteins. *Proc. Natl.*

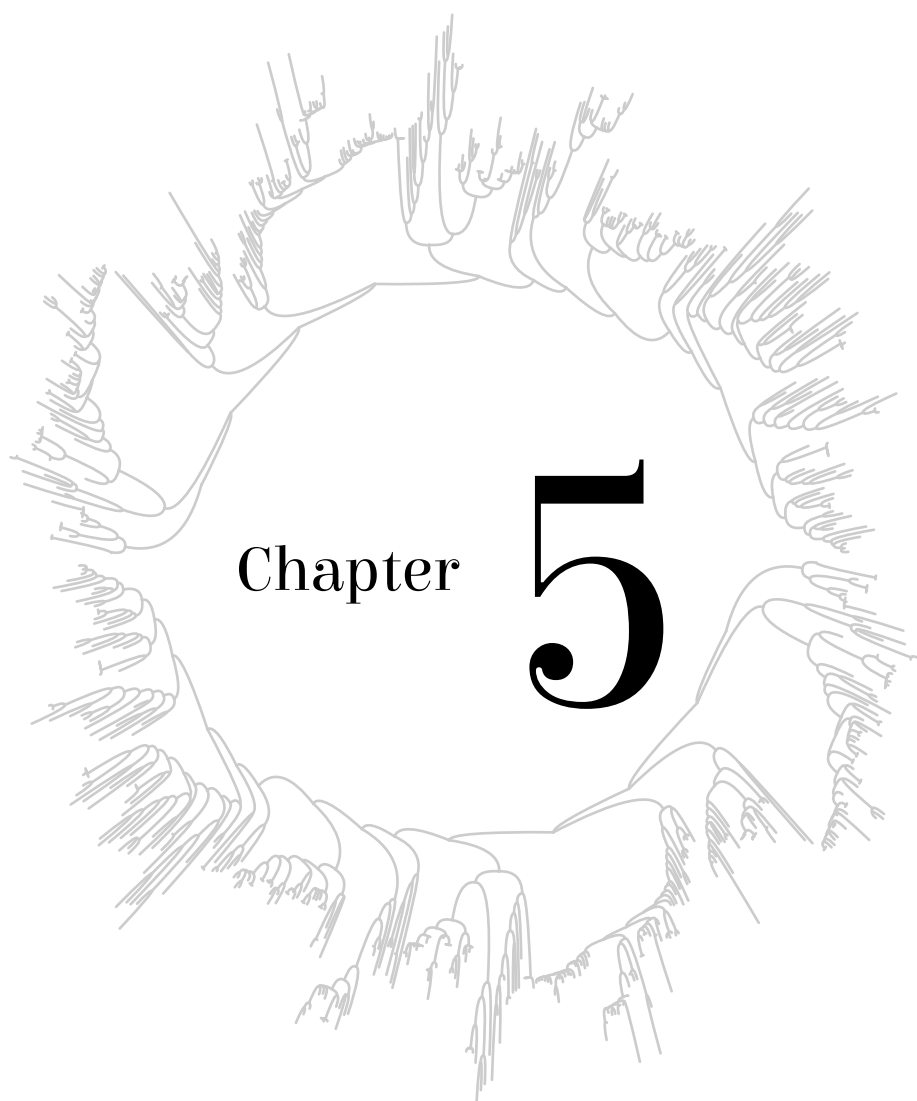


Acad. Sci. USA 94, 11786-11791.

- Korasick, D.A., Westfall, C.S., Lee, S.G., Nanao, M.H., Dumas, R., Hagen, G., Guilfoyle, T.J., Jez, J.M., and Strader, L.C. (2014). Molecular basis for AUXIN RESPONSE FACTOR protein interaction and the control of auxin response repression. *Proc. Natl. Acad. Sci. USA* 111, 5427-5432.
- Kubota, A., Ishizaki, K., Hosaka, M., and Kohchi, T. (2013). Efficient *Agrobacterium*-mediated transformation of the liverwort *Marchantia polymorpha* using regenerating thalli. *Biosci. Biotechnol. Biochem.* 77, 167-172.
- Lanfear, R., Calcott, B., Ho, S.Y., and Guindon, S. (2012). Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29, 1695-1701.
- Lavy, M., Prigge, M.J., Tao, S., Shain, S., Kuo, A., Kirchsteiger, K., and Estelle, M. (2016). Constitutive auxin response in *Physcomitrella* reveals complex interactions between Aux/IAA and ARF proteins. *eLife* 5, e13325.
- Lee, D.J., Park, J.W., Lee, H.W., and Kim, J. (2009). Genome-wide analysis of the auxin-responsive transcriptome downstream of *iaa1* and its expression analysis reveal the diversity and complexity of auxin-regulated gene expression. *J. Exp. Bot.* 60, 3935-3957.
- Liu, X., Huang, J., Wang, Y., Khanna, K., Xie, Z., Owen, H.A., and Zhao, D. (2010). The role of floral organs in carpels, an *Arabidopsis* loss-of-function mutation in *MicroRNA160a*, in organogenesis and the mechanism regulating its expression. *Plant J.* 62, 416-428.
- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550.
- Matasci, N., Hung, L.H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., et al. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience* 3, 17.
- Matías-Hernández, L., Aguilar-Jaramillo, A.E., Marín-González, E., Suárez-López, P., and Pelaz, S. (2014). RAV genes: regulation of floral induction and beyond. *Ann. Bot.* 114, 1459-1470.
- Mallory, A.C., Bartel, D.P., and Bartel, B. (2005). MicroRNA-directed regulation of *Arabidopsis* AUXIN RESPONSE FACTOR17 is essential for proper development and modulates expression of early auxin response genes. *Plant Cell* 17, 1360-1375.
- Monshausen, G.B., Miller, N.D., Murphy, A.S., and Gilroy, S. (2011). Dynamics of auxin-dependent Ca²⁺ and pH signaling in root growth revealed by integrating high-resolution imaging with automated computer vision-based analysis. *Plant J.* 65, 309-318.
- Nanao, M.H., Vinos-Poyo, T., Brunoud, G., Thévenon, E., Mazzoleni, M., Mast, D., Lainé, S., Wang, S., Hagen, G., Li, H., et al. (2014). Structural basis for oligomerization of auxin transcriptional regulators. *Nat. Commun.* 5, 3617.
- Nichols, H.W. (1973). Growth media - freshwater. In *Handbook of Phycological Methods*, J.R. Stein, ed. (London: Cambridge University Press), pp. 7-24.
- Ohtaka, K., Hori, K., Kanno, Y., Seo, M., and Ohta, H. (2017). Primitive Auxin Response without TIR1 and Aux/IAA in the Charophyte Alga *Klebsormidium nitens*. *Plant Physiol.* 174, 1621-1632.
- Paciorek, T., Zajímalová, E., Ruthardt, N., Petrášek, J., Stierhof, Y.D., Kleine-Vehn, J., Morris, D.A., Emans, N., Jürgens, G., Geldner, N., et al. (2005). Auxin inhibits endocytosis and promotes its own efflux from cells. *Nature* 435, 1251-1256.
- Paponov, I.A., Paponov, M., Teale, W., Menges, M., Chakrabortee, S., Murray, J.A., and Palme, K. (2008). Comprehensive transcriptome analysis of auxin responses in *Arabidopsis*. *Mol. Plant* 1, 321-337.
- Piya, S., Shrestha, S.K., Binder, B., Stewart, C.N., Jr., and Hewezi, T. (2014). Protein-protein interaction and gene co-expression maps of ARFs and Aux/IAAs in *Arabidopsis*. *Front. Plant Sci.* 5, 744.
- Plackett, A.R., Rabinowitsch, E.H., and Langdale, J.A. (2015). Protocol: genetic transformation of the fern *Ceratopteris richardii* through microparticle bombardment. *Plant Methods* 11, 37.
- Prigge, M.J., Lavy, M., Ashton, N.W., and Estelle, M. (2010). *Physcomitrella* patens auxin-resistant mutants affect conserved elements of an auxin-signaling pathway. *Curr. Biol.* 20, 1907-1912.
- Rensing, S.A. (2017). Why we need more non-seed plant models. *The New Phytol.* doi: 10.1111/nph.14464.
- Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y., et al. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* 319, 64-69.
- Robert, S., Kleine-Vehn, J., Barbez, E., Sauer, M., Paciorek, T., Baster, P., Vanneste, S., Zhang, J., Simon, S., Covanova, M., et al. (2010). ABP1 mediates auxin inhibition of clathrin-dependent endocytosis in *Arabidopsis*. *Cell* 143, 111-121.
- Saint-Marcoux, D., Proust, H., Dolan, L., and Langdale, J.A. (2015). Identification of reference genes for real-time quantitative PCR experiments in the liverwort *Marchantia polymorpha*. *PLoS one* 10, e0118678.
- Sawa, S., Ohgishi, M., Goda, H., Higuchi, K., Shimada, Y., Yoshida, S., and Koshiba, T. (2002). The HAT2 gene, a member of the HD-Zip gene family, isolated as an auxin inducible gene by DNA microarray screening, affects auxin response in *Arabidopsis*. *Plant J.* 32, 1011-1022.

- Schenck, D., Christian, M., Jones, A., and Lüthen, H. (2010). Rapid auxin-induced cell expansion and gene expression: a four-decade-old question revisited. *Plant Physiol.* 152, 1183-1185.
- Sheard, L.B., Tan, X., Mao, H., Withers, J., Ben-Nissan, G., Hinds, T.R., Kobayashi, Y., Hsu, F.F., Sharon, M., Browse, J., et al. (2010). Jasmonate perception by inositol-phosphate-potentiated COI1-JAZ co-receptor. *Nature* 468, 400-405.
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312-1313.
- Sugano, S.S., Shirakawa, M., Takagi, J., Matsuda, Y., Shimada, T., Hara-Nishimura, I., and Kohchi, T. (2014). CRISPR/Cas9-mediated targeted mutagenesis in the liverwort *Marchantia polymorpha* L. *Plant Cell Physiol.* 55, 475-481.
- Szemenyei, H., Hannon, M., and Long, J.A. (2008). TOPLESS mediates auxin-dependent transcriptional repression during *Arabidopsis* embryogenesis. *Science* 319, 1384-1386.
- Takato, S., Kakei, Y., Mitsui, M., Ishida, Y., Suzuki, M., Yamazaki, C., Hayashi, K.I., Ishii, T., Nakamura, A., Soeno, K., et al. (2017). Auxin signaling through SCFTIR1/AFBs mediates feedback regulation of IAA biosynthesis. *Biosci. Biotechnol. Biochem.* 81, 1320-1326.
- Tan, X., Calderon-Villalobos, L.I., Sharon, M., Zheng, C., Robinson, C.V., Estelle, M., and Zheng, N. (2007). Mechanism of auxin perception by the TIR1 ubiquitin ligase. *Nature* 446, 640-645.
- Ulmasov, T., Hagen, G., and Guilfoyle, T.J. (1999). Activation and repression of transcription by auxin response factors. *Proc. Natl. Acad. Sci. USA* 96, 5844-5849.
- Wang, J.W., Wang, L.J., Mao, Y.B., Cai, W.J., Xue, H.W., and Chen, X.Y. (2005). Control of root cap formation by MicroRNA-targeted auxin response factors in *Arabidopsis*. *Plant Cell* 17, 2204-2216.
- Weijers, D., and Wagner, D. (2016). Transcriptional Responses to the Auxin Hormone. *Annu. Rev. Plant Biol.* 67, 539-574.
- Woodward, A.W., and Bartel, B. (2005). Auxin: regulation, action, and interaction. *Ann. Bot.* 95, 707-735.
- Yang, J., Tian, L., Sun, M.X., Huang, X.Y., Zhu, J., Guan, Y.F., Jia, Q.S., and Yang, Z.N. (2013). AUXIN RESPONSE FACTOR17 is essential for pollen wall pattern formation in *Arabidopsis*. *Plant Physiol.* 162, 720-731.
- Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G., and Bhattacharya, D. (2004). A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biol. Evol.* 21, 809-818.
- Žižková, E., Kubeš, M., Dobrev, P.I., Příbyl, P., Šimura, J., Zahajská, L., Záveská Drábková, L., Novák, O., and Motyka, V. (2017). Control of cytokinin and auxin homeostasis in cyanobacteria and algae. *Ann. Bot.* 119, 151-166.





Deep evolutionary history of the Phox and Bem1 (PB1) domain across eukaryotes

Sumanth Kumar Mutte and Dolf Weijers

Laboratory of Biochemistry, Wageningen University, Wageningen, the Netherlands

Modified version of this chapter has been published as:

Mutte, S. K., & Weijers, D. (2020). Deep Evolutionary History of the Phox and Bem1 (PB1) Domain Across Eukaryotes. *Scientific Reports* 10, 3797. DOI: [10.1038/s41598-020-60733-9](https://doi.org/10.1038/s41598-020-60733-9).



Protein oligomerization is a fundamental process to build complex functional modules. Domains that facilitate the oligomerization process are diverse and widespread in nature across all kingdoms of life. One such domain is the Phox and Bem1 (PB1) domain, which is functionally (relatively) well understood in the animal kingdom. However, beyond animals, neither the origin nor the evolutionary patterns of PB1-containing proteins are understood. While PB1 domain proteins have been found in other kingdoms, including plants, it is unclear how these relate to animal PB1 proteins. To address this question, we utilized large transcriptome datasets along with the proteomes of a broad range of species. We discovered eight PB1 domain-containing protein families in plants, along with three each in Protozoa and Chromista and four families in Fungi. Studying the deep evolutionary history of PB1 domains throughout eukaryotes revealed the presence of at least two, but likely three, ancestral PB1 copies in the Last Eukaryotic Common Ancestor (LECA). These three ancestral copies gave rise to multiple orthologues later in evolution. Analyzing the sequence and secondary structure properties of plant PB1 domains from all the eight families showed their common ubiquitin β -grasp fold, despite poor sequence identity. Tertiary structural models of these plant PB1 families, combined with Random Forest based classification, indicated family-specific differences attributed to the length of PB1 domain and the proportion of β -sheets. Thus, this study not only identifies novel PB1 families, but also provides an evolutionary basis to understand their diverse functional interactions.

Introduction

Phox and Bem1 (PB1) domain: Discovery and properties

Protein-protein interaction is a basic and important mechanism that brings proteins together in a functional module and thus allows the development of higher-order functionalities. One of the versatile interaction domains that brings this modularity through either dimerization or oligomerization is the PB1 domain (Lamark et al., 2003; Noda et al., 2003; Terasawa, 2001). Initially, the two animal proteins, p40Phox and p67Phox, were shown to interact through a novel motif that contains a stretch of negatively charged amino acids (Nakamura et al., 1998). In the same study, it was also shown that the yeast CELL DIVISION CONTROL 24 (Cdc24) protein contains the same motif as found in p40Phox, and hence named as PC motif (for p40Phox and Cdc24; Nakamura et al. 1998). Later, the BUD EMERGENCE 1 (Bem1) protein in yeast was also found to have this motif, after which it has been renamed as PB1 domain (for Phox and Bem1). The PB1 domain of Bem1 in yeast is required for the interaction with Cdc24 to maintain cell polarity (Ito et al., 2001). Later, in mammals, many protein families were identified that contain a PB1 domain (Sumimoto et al., 2007). In plants, the PB1 domain was initially recognised as domain III/IV in the auxin repressor proteins AUXIN/INDOLE-3-ACETIC-ACID (Aux/IAAs) (Hagen and Guilfoyle, 2002). Later, the domains III/IV were found to form a similar fold as (and hence renamed) the PB1 domain, in multiple gene families in plants (Guilfoyle and Hagen, 2012; Svenning et al., 2011; Zientara-Rytter and Sirko, 2014).

The PB1 domain ranges from 80-100 amino acids in length and exhibits a Ubiquitin β -grasp fold with five β -sheets and two α -helices (Korasick et al., 2014; Müller et al., 2006). The first half of the domain represents a positively charged face, with a conserved lysine (K) in β 1. The latter half of the domain represents a negatively charged face, with D-x-D/E-x-D/E as core (OPCA motif; Müller et al. 2006). Based on the presence or absence of these important residues/motifs, the PB1 domains are divided into three types. If the PB1 domain contains only the conserved OPCA motif but not the lysine, it is considered as a type-1 (or type-A) PB1 domain. If there is only lysine but not an OPCA motif, it is a type-2 (or type-B) domain. If the PB1 domain contains both the lysine and the OPCA motifs, it is referred as type-1/2 (or type-AB). Various proteins that harbour a PB1 domain undergo either dimerization or oligomerization, where the positive face of one PB1 domain interacts with the negative face of another in a head-to-tail fashion (Korasick et al., 2014; Lamark et al., 2003). Hence, depending on the type of PB1 domain they interact with, there can be either homotypic or heterotypic PB1 interactions.

PB1 domains in Animalia and Fungi

All the eukaryotes are divided into five kingdoms: Protozoa, Chromista, Fungi, Animalia and Plantae (Adl et al., 2012; Ruggiero et al., 2015). PB1 domain-containing proteins have been relatively well studied in Animalia, when compared to the other kingdoms. At least nine gene families have been shown to encode a PB1 domain (Sumimoto et al., 2007). Animal genomes encode proteins that contain all three types of PB1 domains: type-1 - NEUTROPHIL CYTOSOL



FACTOR 4 (NCF4/p40^{Phox}), MITOGEN-ACTIVATED PROTEIN KINASE KINASE 5 (M2K5) and NEXT TO BRCA 1 (NBR1); type-2 - NEUTROPHIL CYTOSOL FACTOR 2 (NCF2/p67^{Phox}), PARTITIONING DEFECTIVE 6 (Par6) and MITOGEN-ACTIVATED PROTEIN KINASE KINASE KINASE 2/3 (M3K2/3); type-1/2 - SEQUESTOSOME-1 (SQSTM1/p62), ATYPICAL PROTEIN KINASE C (aPKC) and TRK-FUSED GENE (TFG). A systematic analysis through yeast two-hybrid and pull-down assays revealed various homotypic and heterotypic interactions among these PB1 domains (Lamark et al., 2003). The p67^{Phox} upon its interaction with p40^{Phox} activates the phagocyte NADPH oxidase that is important for innate immunity in mammals (Lambeth, 2004). The Par6-aPKC complex establishment through PB1 is essential for cell polarity in mammals and insects (Suzuki, 2006). This complex, along with Par3, also regulates the formation of junctions through apical-basal polarity in mammalian epithelial cells (Joberty et al., 2000). p62 acts as a crucial scaffolding protein playing important roles in autophagy, apoptosis and inflammation (Moscat et al. 2007).

The PB1 domain of M3K2/3 interacts with M2K5 to activate ERK5 mediated signalling in response to growth factors and osmotic stress (Nakamura and Johnson, 2003). TFG PB1 domain is involved in transforming activity by forming the TFG-TrkA (Tyrosine Kinase A) fusion (Roccatto et al., 2003). NBR1 interacts with p62 through PB1 which is required for targeting p62 to sarcomeres (Lamark et al., 2003). Few non-canonical PB1 interactions were also observed, for example, in p40^{Phox} PB1 and PX domains undergo intramolecular interaction, disruption of which is required to activate the NADPH oxidase (Honbou et al., 2007). In yeast, interaction of both the PB1 domain containing proteins, Bem1 and Cdc24 is critical for the cell polarity establishment at both budding and mating (Ito et al., 2001). The NADPH OXIDASE REGULATOR (NoxR) plays a central role in fungal morphogenesis, growth and development through NADPH oxidation pathway (Takemoto et al., 2011).

PB1 domains in Plantae

The best-studied PB1 domains in plants are from the AUXIN RESPONSE FACTOR (ARF) transcription factors and their AUXIN/INDOLE-3-ACETIC-ACID (Aux/IAA) inhibitors. Both the homotypic and heterotypic interactions among and between these gene families is relatively well established (Piya et al., 2014). The structural basis for these interactions has also been scrutinized in detail (Korasick et al., 2014; Nanao et al., 2014). Both ARFs and Aux/IAAs are involved in auxin-dependent gene regulation through the Nuclear Auxin Pathway, that controls various growth and developmental processes (reviewed in Weijers and Wagner 2016). Another PB1 domain containing protein, AtNBR1, an Arabidopsis ortholog of animal NBR1, is involved in autophagy and was shown to homo-polymerize through its PB1 domain (Svenning et al., 2011). Joka2, an AtNBR1 orthologue of tobacco, can also homodimerize through its PB1 domain (Zientara-Rytter and Sirko, 2014). Moreover, this study also revealed non-canonical interaction of the PB1 domain with the C-terminal UBA domain within the same protein (Zientara-Rytter and Sirko, 2014). Homotypic interactions through PB1 domains of NIN-LIKE PROTEINS OF PLANTS (NLPs) are required to induce nitrate-dependent gene expression

(Guan et al., 2017; Konishi and Yanagisawa, 2019). Interestingly, like AtNBR1/Joka2, the PB1 domain of NLP also undergoes non-canonical interaction with the HQ domain of TEOSINTE BRANCHED 1, CYCLOIDEA, PCF DOMAINS CONTAINING PROTEIN 20 (TCP20) (Guan et al., 2017). Another study identified a novel unclassified PB1 domain-containing protein PAL OF QUIRKY (POQ) that undergoes non-canonical interaction with QUIRKY (QKY) (Trehin et al., 2013). However, the structural or mechanistic basis of these non-canonical interactions are yet to be elucidated.

Exploring the unexplored

Even though PB1 domain proteins are well defined and their mechanical basis is relatively well established in animals (reviewed in Sumimoto et al. 2007; Burke and Berk 2015), their evolutionary histories are essentially unknown. Moreover, it is unclear how many PB1 domain-containing gene families are present in other kingdoms. Deep evolution has been relatively well studied for ARF and Aux/IAA gene families (Mutte et al., 2018) and to a certain extent for NLPs (Mu and Luo, 2019), but the presence and the evolution of other PB1 domains, if any, in plants and unicellular eukaryotes is obscure. Hence, the current study is designed to address several important questions related to the distribution and ancestry of PB1 domains in the eukaryotic tree of life: (1) How many PB1 domain-containing gene families are present in the kingdoms Protozoa, Chromista, Fungi and Plantae? (2) What is the origin of the PB1 domain? How many copies of PB1 were present in the Last Eukaryotic Common Ancestor (LECA)? (3) How have PB1-containing proteins diversified/multiplied in evolution across multiple kingdoms? (4) What are the sequence/structural patterns specific to each family of PB1s and how to classify them?

To answer these questions, we have utilized the large transcriptome datasets in Chromista and Plantae and the (almost) complete proteomes from Fungi and Animalia. We found that the PB1 domains have a deep evolutionary origin with at least two copies in LECA. Moreover, we find that the PB1 domain is associated with a variety of domains, ranging from DNA-binding domains to Kinases and membrane-binding domains. Further, a detailed sequence analysis of PB1 domains in Plantae revealed that these are poorly conserved among various families in general, with few residues being specific to each family. Taken together, this study provides the first evolutionary framework of the PB1 domains across the eukaryotes.

Results

Identification and evolution of PB1 domain-containing proteins in various kingdoms

Animalia:

Based on literature, we extracted protein sequences of all PB1 domain-containing proteins in the human genome from the Uniprot database. Nine gene families were found to encode the PB1 domains as a part of their protein architecture (Fig. 1A). aPKC and M3K2/3 both contain PB1 domains as a part of their protein architecture (Fig. 1A). aPKC and M3K2/3 both contain PB1 domains as a part of their protein architecture (Fig. 1A). aPKC and M3K2/3 both contain PB1 domains as a part of their protein architecture (Fig. 1A). Whereas, aPKC contains an extra diacylglycerol-binding (kDAG) domain in the middle. NCF2/p67^{Phox} and NCF4/p40^{Phox} both



contains SRC Homology 3 (SH3) and PB1 domains in the C-terminus, where NCF2 contains Tetratricopeptides and NCF4 contains Phox homologous domain (PX) in their N-terminus (Fig. 1A). The other three protein families, Par6, TFG and p62/SQSTM1, are in general shorter than other PB1 domain-containing proteins, with a PB1 domain in the amino-end. p62 contains a Ubiquitin-associated domain (UBA), whereas Par6 contains a PSD95-Dlg1-Zo1 (PDZ) domain in the carboxy-end (Fig. 1A). The full name or description of all the domains along with a link to the InterPro domain database are provided in Table S1.

PB1 sequences from all the above-mentioned proteins were used as queries to retrieve orthologues from ten species across various phyla in Animalia (see Supplementary file 1 for the list of species used). Retrieved orthologous sequences were used in a phylogenetic analysis along with the respective human counterparts. The PB1 domain-based phylogeny reflected the monophyly of each gene family (Fig. 2 and Fig. S2). PB1 domains of M3K2-M3K3 and aPKC-M2K5 form paralogous pairs, indicating the common ancestry of PB1 for each pair at the emergence of the kingdom Animalia. Interestingly, the paralogous pairs M3K2-M3K3 and aPKC-M2K5 PB1 domains are closer to the respective orthologues from other kingdoms than the other PB1 domains in the same kingdom, Animalia. A similar trend is observed with NBR1, however, surprisingly NCF2/p67^{Phox} is placed as the sister clade to the NBR1. p62 does not show any close relationship with other PB1 domains, neither paralogous nor orthologous, from the same kingdom or the other kingdoms (Fig. 2 and Fig. S2). In a similar way, Par6 and TFG also appear to be Animalia-specific clades (Fig. 2 and Fig. 3).

Protozoa:

From UniProt, at least six reference proteomes and other individual Protozoan sequences from various species across different phyla in Protozoa were used to identify the PB1 domains (Supplementary file 1). Few PB1 domain-containing proteins were identified along with a large number of partial (or truncated) proteins with either only a PB1 domain or a large unknown flanking sequence (Fig. 1 and Fig. S1). Among the (full length) PB1 domain-containing proteins, orthologs of Animalia M3K2/3 as well as Plantae Phox were identified, and named as Kinase and Phox respectively (Fig. 1B and Fig. S2). Unlike the animals, the Protozoan kinases also contain WD40 repeats at their C-terminus. Moreover, the PB1 domain is also adjacent to kinase domain (Fig. 1B). Orthologues of the NBR1 with all the four (known) domains were also found, along with the sequences of various domain combinations i.e. either PB1 with Zn-finger, with the NBR1 Central domain, or with the Ubiquitin Associated (UBA) domains (Fig. 1). We also identified many PB1 domain-containing proteins either associated with a Sterile Alpha Motif (SAM), Per-Arnt-Sim (PAS), EF-hand or Cystathionine Beta Synthase (CBS) domains (Fig. S1). However, the majority of these are identified in only one sequence or one species, and also represented in polyphyletic groups spread across the phylogenetic tree, making it difficult to classify them into a certain clade (Fig. S2).

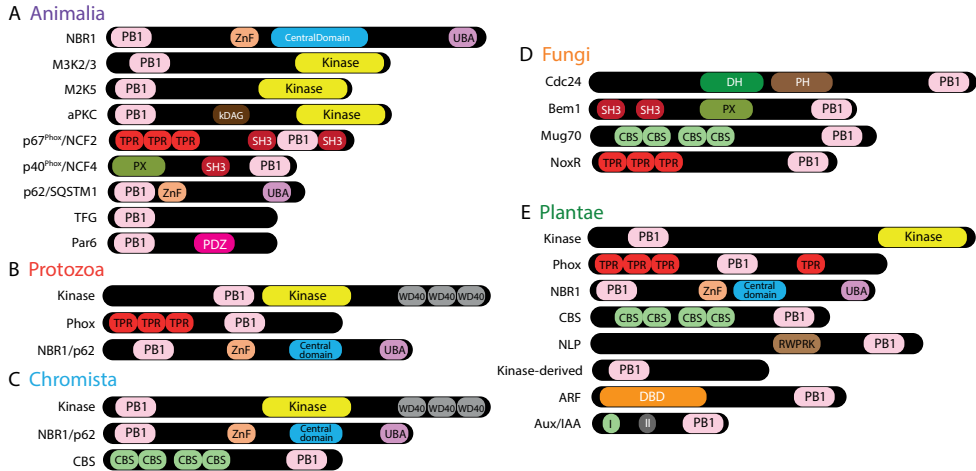


Figure 1: Domain architecture of various PB1 domain containing proteins across the five eukaryotic kingdoms. DBD in (E) represents the DNA binding domain, which is a combination of both B3 and dimerization domain (DD). Abbreviation and the corresponding InterPro database link of all the domains are provided in Table S1. PB1 domains that are identified in only one sequence and/or one species are provided in the Figure S1.

Chromista:

To identify the PB1 domains in Chromista, all the transcriptomes from the MMETSP database were used (Supplementary file 1). Well-annotated fungal (Bem1 and Cdc24), plant (*Arabidopsis* and *Marchantia*) and animal (Human and Mouse) PB1 sequences were used to query the database, and processed the data with the pipeline developed earlier (Chapter 2; Mutte et al., 2018). Two gene families, Kinase and NBR1 were identified in Chromista, with a similar domain architecture, being orthologous to the respective gene families in Animalia (Fig. 1C and Fig. S2). Interestingly, like Protozoan kinase proteins, these carry WD40 repeats too, however, the PB1 domain is far to the N-terminus as in Plantae or Animalia (Fig. 1C). Orthologues of NBR1 are identified as multiple (partial) proteins represented as polyphyletic groups, similar to Protozoa (Fig. S2). However, as a third gene family, we identified the CBS domain-containing proteins, where the PB1 domain is associated in their carboxy terminus (Fig. 1C). Few other PB1 domain proteins were identified as single copies in only one species that host either a Tetratricopeptide (TPR) repeat or an EF-hand domain, which were also represented in polyphyletic groups (Fig. S1 and Fig. S2).

Fungi:

For Fungi, we have selected the 12 reference proteomes from MycoCosm database (Supplementary file 1). Well-annotated plant and animal PB1 sequences were used as query sequences. Four PB1 domain containing protein families were identified (Fig. 1D). The widely known Bem1 and Cdc24, were identified as a monophyletic paralogous pair in our study (Fig. 2). Along with the PB1 domain, Bem1 contains SH3 and PX domains, whereas Cdc24 contains Dbl homology

(DH) and Plextrin homology (PH) domains. Interestingly, NCF4/p40Phox, the animal ortholog of Bem1, contains PX in N-terminus, unlike in the middle as in Bem1 (Fig. 1D). CBS domain containing proteins (referred as Mug70) were also identified, with a similar domain architecture like in other kingdoms (Fig. 1D). Further, NoxR, an ortholog of Animalia and Plantae Phox, was also identified as a sister clad to this pair (Fig. 2 and Fig. S2).

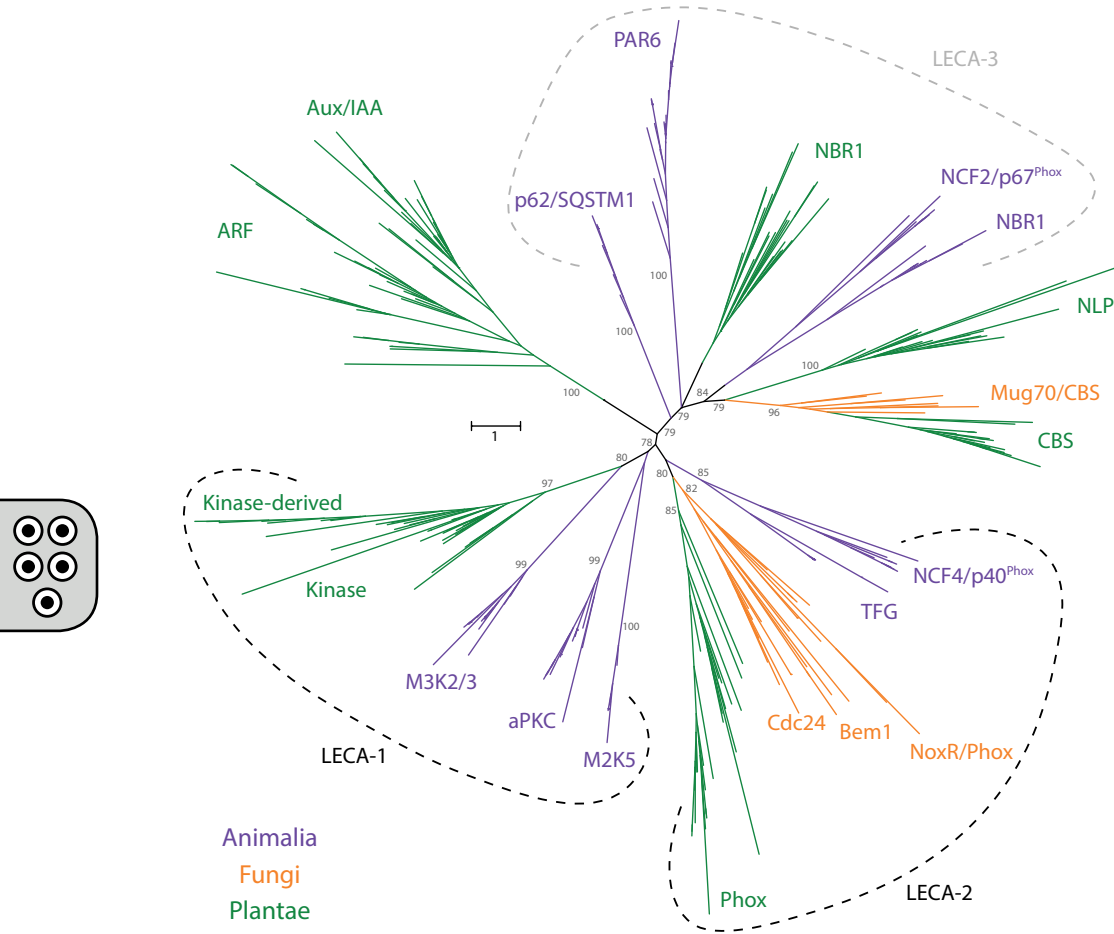


Figure 2: Unrooted tree with representative Fungi, Animalia and Plantae PB1 domains. Early branches that are well-supported (bootstrap >75) are indicated in grey. Orthologs from each kingdom are represented with each colour as indicated: Fungi in ‘orange’, Animalia in ‘purple’ and Plantae in ‘green’. The groups outlined with dotted lines indicated as LECA-1, LECA-2 and LECA-3 represent the probable ancestral copies in LECA corresponding to Kinase, Phox and NBR1 groups respectively. Another phylogenetic tree with all the five kingdoms is presented in the Figure S2 as schematic and full version with taxa names and domain information of both the trees can be found at iTOL: <https://itol.embl.de/shared/dolfweijers>.

In summary, all the four gene families form a respective individual monophyletic group with all the paralogs, indicating their presence across major phyla in Fungi (Fig. 2 and Fig. 3). It is worth noting that an ortholog of p62 and a PB1 domain associated with a SAM domain were identified. However, each was found in only one species and a single copy (Fig. S1). Hence,

we discarded them for further analysis as they are considered of low confidence and may not represent any phylum or the kingdom itself.

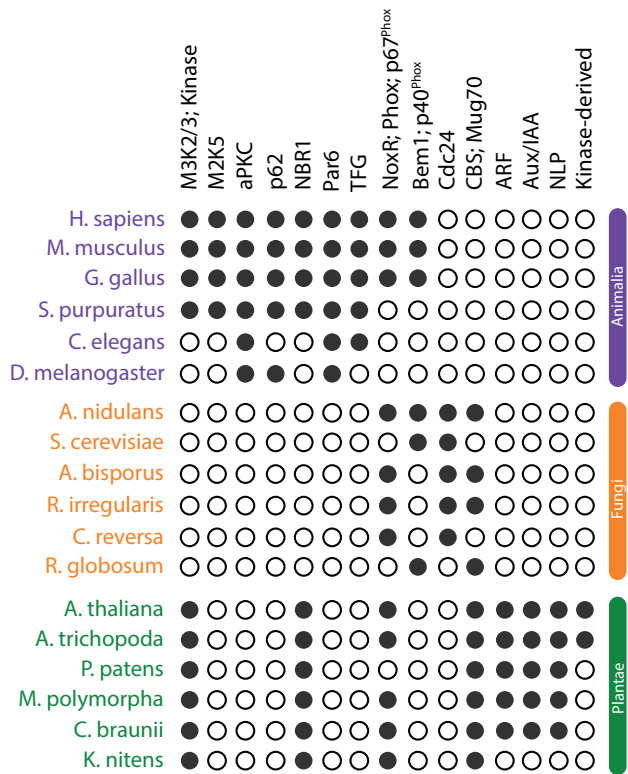


Figure 3: Summary of various PB1 domain-containing proteins across key species in Animalia, Fungi and Plantae. Filled and empty circles represent presence and absence, respectively, of the orthologous genes for the family mentioned on the top in the corresponding species.

Plantae:

To identify all the PB1 domains in the kingdom Plantae (Fig. 1E), we have adapted a similar pipeline as mentioned above (Chapter 2; Mutte et al., 2018), using 485 transcriptomes, that belong to multiple phyla in the kingdom Plantae, from the OneKP database (Carpenter et al., 2019; Supplementary file 1). We identified eight gene families that encode for PB1 domain-containing proteins in plants (Fig. 1E). Among these, NBR1 and Kinase orthologues are placed in the same clade as their counterparts from other kingdoms, and also contain the same domain architecture as their animal orthologs (Fig. 1E and Fig. 2). ARF and Aux/IAA families form a distinct monophyletic clade indicating a common ancestry at the base of the streptophytes. ARFs contain B3 and dimerization domains (together referred as DNA-binding domain (DBD)) at the N-terminus and a PB1 domain at the C-terminus, like the Aux/IAAs. In addition to a PB1 domain, Aux/IAAs also contain an EAR-motif and a degron motif (domain I and II respectively; Fig. 1E). Phox proteins, having the same domains as animal counterparts, form a sister clade

to the respective orthologous proteins from other kingdoms (Fig. 2). CBS domain-containing proteins were also identified in plants, placed in the same clade as fungal Mug70. Kinase-derived, ARF, Aux/IAA and NLP are Plantae specific families that are not identified in any other kingdom (Fig. 3). All these plant-specific gene families were discovered before, except Kinase-derived, which has only a PB1 domain in its N-terminus with a large flanking sequence without any known domains. It is worth mentioning that the Kinase-derived PB1 domains, resemble the Kinase PB1 domains and appears to have been duplicated in the ancestors of angiosperms (Fig. 2 and Fig. 3). NLPs contain an RWP-RK domain, in association with the PB1 domain and they are placed as sister clade to the CBS domain containing proteins (Fig. 1E and Fig. 2). Interestingly, among all the families identified so far across all the kingdoms, there do not appear to be any constraints on either the position of the PB1 domain in the protein, or the category of domains it is associated with (DNA binding, oligomerization, phosphorylation etc.; Fig. 1). An overview of all the identified gene families and their existence across the major species in the kingdoms Animalia, Fungi and Plantae is summarized in Fig. 3.

Ancestral copy number in LECA

To better understand the origin and evolutionary patterns of all the PB1 domains across the five kingdoms in eukaryotes, two phylogenetic trees were constructed using only the PB1 domain protein sequences. One is based on the PB1 domains from only three kingdoms (Animalia, Fungi and Plantae; Fig. 2), whereas another one is constructed based on all the sequences from five kingdoms (Fig. S2). The detailed versions of both the phylogenetic trees are available online in the iTOL webserver: <https://itol.embl.de/shared/dolfweijers>. All the previously mentioned pairs that form the monophyletic groups of individual families in each kingdom are well supported with good bootstrap values (>75), especially in Fungi, Animalia and Plantae (Fig. 2). The branches representing PB1 domains in Protozoa and Chromista are highly unreliable due to the polyphyletic nature and their random distribution across the phylogenetic tree (Fig. S2). Overall, the recently evolved clades in the phylogeny that are either gene family-specific or kingdom-specific, are generally monophyletic in nature. We have observed a decrease in the support of the split of early branches (with poor bootstraps) in the phylogeny based on all the five kingdoms (Fig. S2; refer iTOL tree). Monophyletic grouping, as well as the presence in multiple kingdoms, support the notion that there would have been at least two common ancestral copies of PB1 domains, each corresponding to Kinase and Phox orthologues across eukaryotes. Even though the Plantae NBR1 PB1 domains, along with the animal orthologues (and the similar proteins p62) are not monophyletic in origin, they are still placed in the phylogeny as sister clades (Fig. 2). This distribution of orthologues from the various kingdoms hint at a third common ancestor of PB1 in LECA. This analysis has failed to predict the order of evolution because of the lack of sufficient phylogenetic signal due to poorly conserved sequences, a relatively small domain (in general) and poor bootstraps in the early branches in the tree with all five kingdoms. The use of bacterial outgroup sequences could not improve resolution, leading to mixing in the phylogeny along with ingroup sequences. Hence, no outgroup was used and the tree is unrooted.

Because of these drawbacks, this study could not identify the order of events, but could predict the copy number in LECA, based on both the monophyletic nature of Kinase and Phox groups as well as presence of NBR1 orthologous sister clades across multiple kingdoms.

(Dis)similarities in the plant PB1 domains

After identifying the majority, if not all, of the PB1 domain-containing proteins and understanding their evolution patterns across major phyla in all five kingdoms in eukaryotes, we further investigated the plant PB1 domains in detail at the amino acid level. To achieve this, we gathered the PB1 domain sequences from four whole genome-sequenced land plants, one species each from liverworts (*Marchantia polymorpha*), mosses (*Physcomitrella patens*), basal angiosperms (*Amborella trichopoda*) and a core eudicot (*Arabidopsis thaliana*). All PB1 domain protein sequences that belongs to the eight families identified were aligned, and an individual sequence logo was derived for each family (Fig. 4). The well-conserved (group of) residues across the majority of the families are the positive residues lysine (K) in β 1 and arginine (R) in β 2 that together represent the positive surface. However, the lysine of β 1 that makes contact with the OPCA motif on the negative face of another PB1, is not conserved in Kinase and Kinase-derived PB1 domains, indicating that these could be type-1 PB1 domains with only a conserved negative face (Fig. 4).

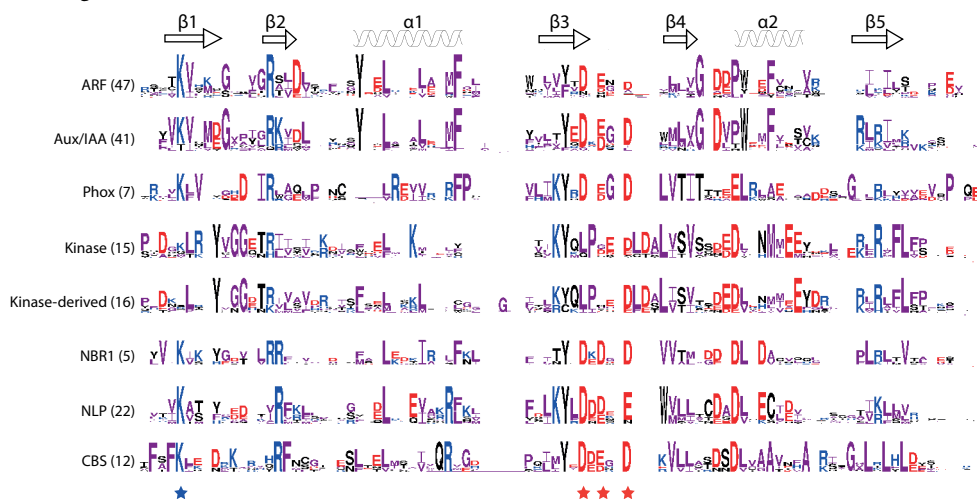


Figure 4: Sequence logos based on the alignment of PB1 domains from the representative land plants (*Marchantia*, *Physco*, *Amborella* and *Arabidopsis*). Secondary structures (α -helices and β -sheets) represented on the top are based on the ARF5 structure (PDB ID: 4CHK). Numbers represented in the braces next to the name of the gene family, shows the number of sequences present in all these four species together, and also the number of sequences used for that particular alignment logo. Amino acids are coloured according to the group: 'PAGFLIMV', 'KRH' and 'DE' are shown in 'purple', 'blue' and 'red' colours respectively. All other amino acids are shown in 'black'. Stars at the bottom represent the key residues on positive (blue) and negative (red) faces, corresponding to Lysine and OPCA motif (D-x-D/E-x-D/E core) respectively.

In general, the negative face represented by the OPCA motif is relatively well-conserved in all the gene families, despite a strong conservation of three amino acids (QLP) just before the

OPCA motif in Kinase and Kinase-derived PB1 domains (Fig. 4). Interestingly, the tyrosine (Y) in $\beta 3$ is relatively well conserved in all the gene families. Apart from these generally conserved residues across multiple families, there are various single amino acids that are specifically conserved in each gene family. For example, tyrosine (Y) and phenylalanine (F) of the $\alpha 1$, glycine (G) in $\beta 4$ and phenylalanine (F) in $\alpha 2$ are specific to ARF and Aux/IAA PB1 domains. In a similar way, two phenylalanine (F) in and before the $\beta 1$ and a G-x-L-x-L-x-L motif in $\beta 5$ are specific to PB1 domains associated with CBS domains (Fig. 4). The tryptophan (W) in $\beta 4$ is specific to NLPs. Despite NBR1 being a single-copy gene in the kingdom Plantae, there do not seem to be any constraints on the domain itself, as there is less than 20% identity among them. This provides a basic understanding of relaxed evolutionary pressure in the PB1 domain, providing opportunities for many gene family-specific changes. This makes it difficult not only to predict general sequence patterns that are important for function, but also to estimate the domain properties specific to each family purely based on the primary sequence and its poorly conserved amino acids.

Classification using Random Forests

Since there are no clear patterns to identify the gene family to which each PB1 belongs to and because it is also not possible to identify important features of a specific PB1 domain based on the sequence alignment, one might detect patterns based on the secondary structure composition along with the amino acid properties. Random forest (RF) based classification was performed with 28 amino acid descriptors as variables. After bootstrap aggregating (bagging) all the decision trees from the RF, the mean out-of-bag (OOB) error rate is only 6% which indicates the high reliability of the RF model (Fig. 5). The classification error rate is the highest (~14%) for Kinase-derived and the least (~2%) for ARF PB1 domains (Fig. 5A and Table S3). On an average, the majority of PB1 families were resolved well, indicating the high reliability of classification using these descriptors.

The importance of each variable is evaluated through the mean decrease in accuracy (MDA) and the mean decrease in Gini (MDG). Higher values of both MDA and MDG indicate the most important variables. In this case, the top 10 important variables are shown in Fig. 5B, with mHbeta being the most important variable to differentiate the different classes of PB1 domains. Hydrophobic moment of β -sheets, mHbeta, indicates the strength of periodicity in the hydrophobicity of the β -sheets, also indicating the formation of more β -sheets (Eisenberg 1984). The next most important variables are composition of proline (P) and length of the PB1 domain (Fig. 5B). We further analysed how these three important variables differ between the gene families (Fig. 5C). mHbeta is low for ARFs and Aux/IAAs, slightly higher for Phox, but even higher for the rest of the gene families. On the other hand, the composition of proline is lowest in CBS, but shows a very broad distribution in the Kinase-derived family. However, the length of the PB1 domain is very constrained for majority of the families (>90 for NBR1 and <90 for ARFs), except Aux/IAA and Kinase-derived, and to a certain extent for Phox (Fig. 5C). To correlate the contribution of mHbeta to the β -sheets in the secondary structure, we

have performed homology modelling of at least one randomly selected *Arabidopsis* orthologue and we indeed found that higher mHbeta represents secondary structures with more β -sheets (Fig. 6). For example, IAA17 shows ~18% of the residues in β -sheets, whereas CBS36500 has ~34%, correlating with lower and higher mHbeta values observed for Aux/IAA and CBS PB1's, respectively (Fig. 6 and Fig. S3). Taken together, these results clearly indicate that there is a difference between the gene families that can be explained from the mHbeta, the composition of proline and by keeping the length unique/constrained for that respective family.

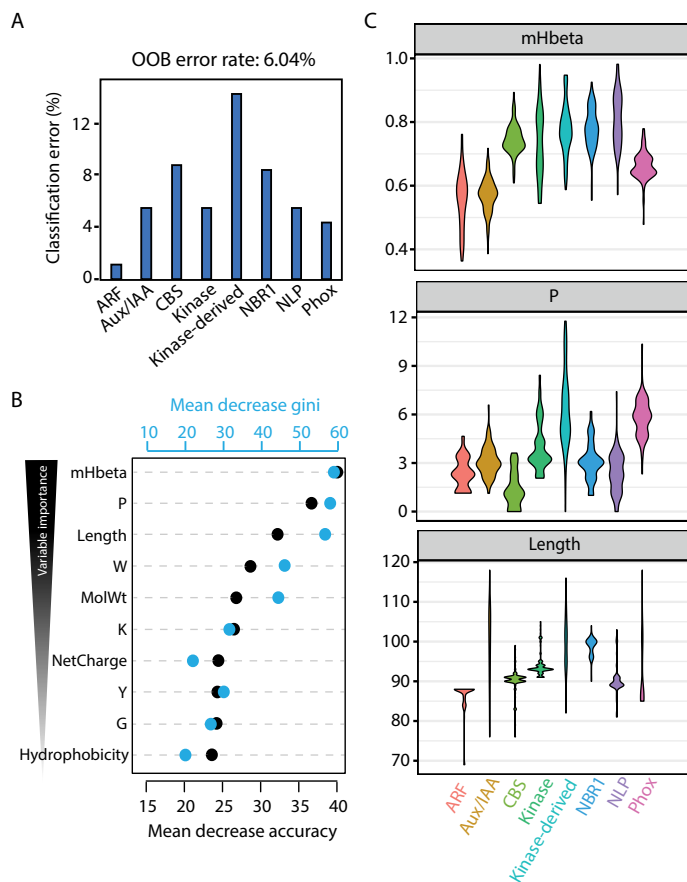


Figure 5: Random forest (RF) classification of the plant PB1 families. (A) Mean out-of-bag (OOB) error rate of 6% is reported for the classification of eight plant PB1 families with individual classification error % as shown in bar chart (B) Importance plots of 10 most important descriptors (variables). The predictive value of each variable was expressed as the mean decrease in accuracy (black dots with scale at bottom) and the mean decrease in Gini (blue dots at top), arranged from most important (top) to the less important (bottom) variables. (C) Violin plots showing the actual distribution of three most important variables across eight families. mHbeta has no units; Composition of Proline is indicated as percentage; and Length is the number of amino acids in the PB1 domain.

Discussion

The PB1 domain is widespread in nature, throughout all the kingdoms in the eukaryotic tree of life. It is diversified to a great extent in organisms with complex body plans like animals and even more in land plants (Fig. 1). As a result, the human genome encodes 13 PB1 domain-containing proteins, whereas a simple model angiosperm, *Arabidopsis thaliana*, encodes more than 80 PB1 domain copies grouped into eight families (Fig. 3). The proteins with a PB1 domain also feature various domains, representing a manifold association ranging from DNA/protein binding, catalytic function, scaffolding to membrane association (Fig. 1). However, the PB1 domain is (mostly) found at either terminus of the protein, preferably facilitating these to perform their native function, scaffolding or oligomerization, without the hindrance of other domains.

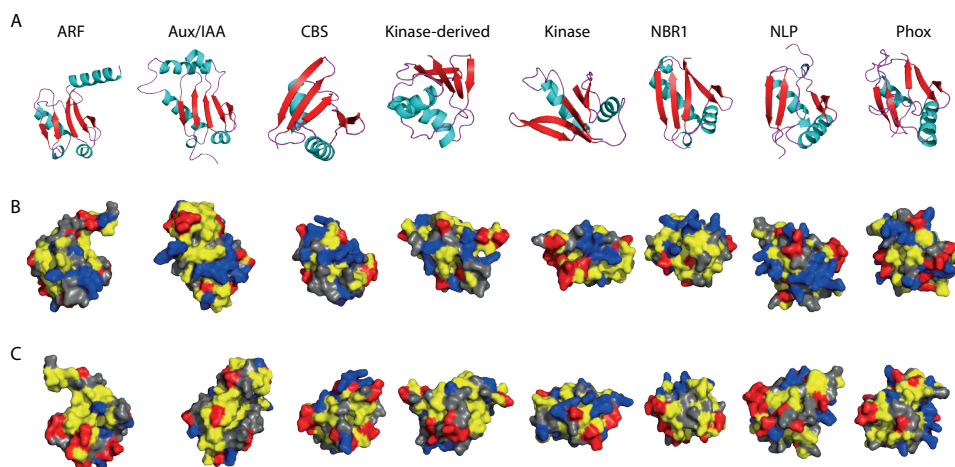


Figure 6: Representative homology model structures of one member from each family in Arabidopsis. The identifiers were: ARF5 (PDB: 4CHK), IAA17 (PDB: 2MUK), At2G36500 (CBS), At1G04700 (Kinase), At2G01190 (KinaseDerived), NBR1, NLP9 and Phox2. (A) Secondary structures shown in various colours: α -helices in 'Cyan'; β -sheets in 'red' and turns in 'purple'. Surface representation of the positive and negative faces shown in (B) and (C) respectively. Hydrophobic amino acids 'AGVILFMP' are in 'Yellow'; Polar residues 'NQTSYCW' are in 'Grey'; Positively charged 'RKH' are in 'Blue' and Negatively charged 'DE' are in 'Red'.

The evolutionary patterns of the PB1 domain showed that there are multiple families shared across multiple kingdoms (Fig. 2). Kinase and NBR1 are present in all the five kingdoms, while Phox is found in four kingdoms (except Chromista) with a similar domain architecture (Fig. 1). The phylogenetic placement of Kinase and Phox PB1 domains and their orthologs indicate the presence of two ancestral copies in LECA and presumably a third copy might be represented by the NBR1 and/or p62 group. It is known that orthologous proteins may perform similar functions by interacting with similar proteins across kingdoms. An example of common functionality of orthologous domains across multiple kingdoms is the recently studied DIX domain in plant cell polarity protein SOSEKI (van Dop et al., 2020). The DIX head-to-tail oligomerization domain is conserved across multiple kingdoms (e.g. DISHEVELLED in

animals; DIX-like in protozoans), and functions in cell polarity by forming oligomers in plants and animals. In a similar way, the PB1 orthologs across multiple kingdoms may share common functionality in similar pathways. One such is the PB1 domain containing protein NBR1, which serves as an autophagy cargo receptor in both plants and animals, where homodimerization of the PB1 domain is also conserved as a part of its function (Kirkin et al., 2009; Svenning et al., 2011). Phox orthologues in animals (p67Phox) and fungi (NoxR) play a key role in NADPH oxidation pathway by interacting with the membrane associated proteins (gp91^{Phox} and NoxA/B respectively) as well as through PB1 domain with p40^{Phox} and Cdc24 respectively (Sumimoto, 2008; Takemoto et al., 2011). Similarly, in *Arabidopsis*, Phox4 was shown to interact with membrane-associated proteins KNOLLE, SYP22 and PEN1, which belong to the SNARE family (Fujiwara et al., 2014). However, it is unknown which PB1 domain protein interacts with Phox4 PB1 in plants. In another study, Phox proteins (referred as MadB) were shown to be involved in the myosin-driven interactions, preferably through PB1 domain (Kurth et al., 2017). Hence, discovering these unknown and novel interactions may provide a link to the existence of common pathways in plants controlled by PB1-dependent interactions.

Apart from the proteins that are shared across multiple kingdoms, some are specific to each kingdom (Fig. 3). ARF, Aux/IAA and NLP families are specific to plants, whereas TFG, Par6, M2K5 and aPKC are specific to animals. ARFs and Aux/IAAs are involved the Nuclear Auxin Pathway, controlling transcriptional regulation of downstream targets with multiple functions in response to the phytohormone auxin (reviewed in Weijers and Wagner 2016). NLPs are master regulators of nitrate-inducible gene regulation in higher plants (Konishi and Yanagisawa, 2019). On the other hand, animal Par6 and aPKC PB1 domains are known to interact with each other playing a key role in cell polarity (Hirano et al., 2005). Thus, as far as can be inferred, these kingdom-specific PB1 domain-containing proteins appear to regulate processes that are specific to that kingdom.

It is interesting to see that the key interacting PB1 domains have also evolved in pairs. Some such pairs are: aPKC-M2K5 (Animalia), Bem1-Cdc24 (Yeast), ARF-Aux/IAA (Plantae) (Fig. 2). The interacting pairs (for example ARF-Aux/IAA) seem to maintain pairs of amino acids specific to those classes (Fig. 4). Hence, based on this 'paired' conservation pattern, it is enticing to speculate that the Kinase and Kinase-derived PB1 domains might form interacting pairs (Fig. 4). Despite the overall poor sequence conservation, it is clear that PB1 domains are maintaining a flexible (global β -grasp fold) yet specific (local conserved residues) sequence context in each family may provide specificity in function. Adding to the complexity in specificity of each interaction, the PB1 domains can also undergo non-canonical interactions. In plants, PAL OF QUIRKY (POQ), a Kinase-derived PB1 domain, interacts with QUIRKY (Trehin et al., 2013). The PB1 domain of NLP interacts with HQ domain of TCP20 (Guan et al., 2017). In animals, the M2K5 PB1 interacts with ERK5, among many others (Sumimoto et al., 2007). However, the structural and/or mechanistic basis of any of these interactions is currently unknown.

In various kingdoms, the PB1 domain-containing proteins have expanded to various



complexities/copies. For example, NBR1 in plants is (mostly) a single copy gene, where ARFs and Aux/IAAs are represented by large gene families with more than 20 copies (Table S2). This clearly shows varying duplication rates in different gene families. However, whether it is a single- or a multi-copy gene family, there is hardly any conservation in the PB1 domain among the members of the same gene family outside of key residues: lysine in β 1, tyrosine in β 3 and the OPCA motif (Fig. 4). Despite their low conservation, all the PB1 domain families identified in plants can potentially form a β -grasp ubiquitin fold (Fig. 6). Thus, for the PB1 domain it is evident that sequence conservation seems to be a less important factor than maintaining the overall β -grasp structure itself.

This poor sequence conservation is never a bottleneck to identify the most important features, as there are efficient machine learning based classification programs like Random Forests (RF). RF has been very successful in classification with highly correlated variables at low error rate (Breiman, 2001). The classification error rate is as low as 2% (in ARFs), but up to 14% in Kinase-derived, which could be due to the broader distribution of all three most important variables (Fig. 5C). This clearly defines that the more specific the variables are, the lower the error rate is. RF also provides the relative importance of each variable with the precision. Hydrophobic moment of β -sheets, mHbeta, the most important variable in our case, is low for Aux/IAA but high for CBS, correlating with the increased β -sheets in CBS (Fig. 6). How this increase of β -sheets could bring a change in function needs to be elucidated. Another interesting observation is that there is a clear difference between some variables being very constrained for each family. For example, the length of the PB1 domain is always above 90 amino acids in the NBR1 family, where as it is always below 90 for ARFs. Hence, it is evident that PB1 domains are constrained in different ways to maintain the uniqueness of that family. Moreover, using more (specific) parameters in the future, one should be able to distinguish PB1 domains to a much broader extent, even across multiple kingdoms, and including homotypic and heterotypic interactions.

Apart from DIX and PB1 domains, the SAM domain also undergoes head-to-tail oligomerization, but this domain is structurally different from both others (Bienz, 2014). It is unclear why the PB1 domains are much more widespread compared to DIX or SAM domains. The latter are only limited to few families and few members in each family. One reason could be that, as discussed above, the PB1 domain might contribute additionally by a wide range of non-canonical interactions and its abundance across multiple kingdoms.

Materials and Methods

Search for PB1 domains in Animalia, Protozoa and Fungi

To study the PB1 domains in the kingdom Animalia, based on the literature, we first extracted all the PB1 domain sequences of Human proteome from the UniProt database (<https://www.uniprot.org/proteomes/>). To find other PB1 domains, we then used ten proteomes from the kingdom across various phyla (Supplementary file 1). A protein database has been created with all these proteomes and queried this database with the PB1 domain sequences from already

known plant (*Arabidopsis* and *Marchantia*) and animal (Human) species. BLASTP module in NCBI BLAST 2.7.1+ (Camacho et al., 2009) was employed for this search and InterPro domain database v5.30-69.0 (<https://www.ebi.ac.uk/interpro/>) was used for domain identification in the BLAST hits. All the sequences that have a PB1 domain have been used for further phylogenetic analysis. A similar procedure was used to obtain the PB1 sequences from Protozoa and Fungi. However, the proteomes of twelve fungi across multiple phyla have been obtained from MycoCosm database at JGI (<https://mycocosm.jgi.doe.gov>). For the Protozoa, we have used the six reference proteomes from UniProt (Supplementary file 1).

Identification of the PB1 domains in Chromista and Plantae

To identify the PB1 domains in the kingdom Plantae, we employed a large transcriptome resource, 1000 plant transcriptomes (OneKP) database (Matasci et al. 2014; www.onekp.com). Out of nearly 1300 transcriptomes in the database, we have used 485 transcriptomes in this study, covering all the phyla in the kingdom Plantae. We have adapted a protocol that was developed earlier (Chapter 2; Mutte et al., 2018). In brief, the query PB1 sequences from *Arabidopsis* were searched against each transcriptome, where the resulting scaffold hits were translated using TransDecoder (v2.0.1; <https://transdecoder.github.io>). All these translated sequences were checked for the presence of PB1 domains using InterProScan (Jones et al., 2014) and only those protein sequences with a PB1 domain identified were used for further analysis. In a similar way, for Chromista, we have employed another transcriptome dataset, Marine Micro Eukaryote Transcriptome Sequencing Project (MMETSP) database (Keeling et al., 2014). We have used all the available transcriptomes and adapted a similar protocol as mentioned above (Supplementary file 1).

Phylogeny construction and visualization

Using all the PB1 sequences that were identified in all the five kingdoms of eukaryotes, we performed the phylogenetic analysis (Supplementary file 2). The protein sequences were aligned with MAFFT G-INS-i algorithm using default parameters (v7; Katoh and Standley 2013). Alignment was cleaned up further, where the positions with more than 20% gaps were removed with trimAl, prior to phylogeny construction (Capella-Gutiérrez et al., 2009). ModelFinder (accessed through IQ-TREE) indicated 'LG' as the best model of evolution, of all the 462 models tested (Kalyanamoorthy et al., 2017). Further, the Maximum Likelihood (ML) method, employed in the IQ-TREE program was used for the phylogenetic tree construction, with 1000 rapid bootstrap replicates and tree branches tested by SH-aLRT method (Nguyen et al., 2015). The resulting tree was manually curated further for some misplaced taxa. An unrooted tree was visualized in iTOL v4 (<https://itol.embl.de/shared/dolfweijers>). In a similar way, another phylogenetic tree was generated using the PB1 sequences only from three kingdoms (Animalia, Fungi and Plantae).



Alignment of the plant PB1 domains

To understand the PB1 domains in the plant kingdom further, we have taken the PB1 sequences of all the eight families from four species (*Marchantia*, *Physcomitrella*, *Amborella* and *Arabidopsis*), aligned them using ClustalOmega (Sievers and Higgins, 2018). After the alignment, the domains from each family were separated and a sequence logo was generated using All the gene identifiers from these four species are available in Table S2. LogOddsLogo server was used for logo generation, with the colour codes for specific amino acids (<https://www.ncbi.nlm.nih.gov/CBBresearch/Yu/logoddslogo/proteins.cgi>). Amino acid groups ‘PAGFLIMV’, ‘KRH’ and ‘DE’ were shown in purple, blue and red colours respectively. All other amino acids were shown in black colour.

Random forest (RF) based plant PB1 classification

The Random Forest (RF) method was used to identify key amino acid descriptors to differentiate and classify each of the PB1 domains into eight plant PB1 families (Breiman, 2001). To make this classification an extensive one, we extracted all the PB1 domains from Plaza Monocots database v4.5, that includes genomes from all the major phyla in Embryophytes (Van Bel et al., 2018). Since the size of each family is different, and to make the analysis uniform and comparable, we have extracted 100 PB1 sequences randomly for each gene family (except 78 for NBR1 as it is a single copy gene). We have used 28 amino acid descriptors (variables) calculated either with ‘protr’ or ‘peptides’ R packages (Osorio et al., 2015; Xiao et al., 2015). Among these, 20 variables correspond to the composition of 20 amino acids, and the remaining eight correspond to the general parameters such as length, molecular weight, hydrophobicity, net charge, isoelectric point (pI), aliphatic index, hydrophobic moment of alpha and beta sheets. We used ‘RandomForest’ R package to build a maximum of 500 decision trees with 5 variables being tried at each step (www.r-project.org; Liaw and Wiener 2002). Confusion matrix and variable importance plots showing mean decrease in accuracy and gini were obtained. Descriptive plots and other graphs shown were obtained using ‘ggplot2’ R package. Supplementary file 3 provides the complete R script that has been used for the RF analysis.

Homology modelling

Homology modelling for the eight selected members in *Arabidopsis*, one each from each plant PB1 family were performed using Phyre2 webserver ‘normal’ mode (Kelley et al., 2015). The identifiers of the PB1 sequences used were: ARF5 (PDB: 4CHK), IAA17 (PDB: 2MUK), At2G36500 (CBS), At1G04700 (Kinase), At2G01190 (Kinase-derived), NBR1, NLP9 and Phox2. Obtained homology models were visualized in PyMol software (Schrodinger Inc., USA).

Supplementary files

All the supplementary files are available online under the URL: <https://www.nature.com/articles/s41598-020-60733-9>.

Acknowledgement

We thank Mark Roosen, Dr. Francois Parcy and Dr. Mariann Bienz for helpful comments on the manuscript. This research was supported by a VICI grant to D.W., from the Netherlands Organization for Scientific Research (NWO; 865.14.001).

Supplementary information

Supplementary file 1: Excel sheet showing the lists of species used in all five kingdoms

Supplementary file 2: FASTA file with the PB1 sequences used for phylogeny

Supplementary file 3: R script used for the Random Forest (RF) classification and descriptive statistics

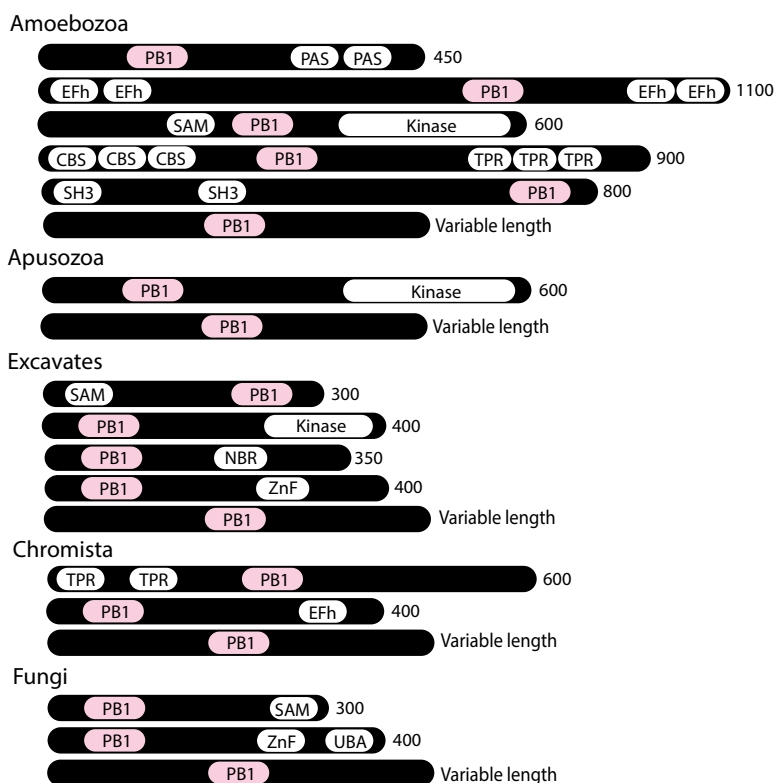


Figure S1: Presence of various PB1 domain containing proteins identified in only one sequence and/or one species. The numbers at the end of each row is the approximate length of the protein. The complete information about the domains and their respective InterPro database links are provided in Table S1.

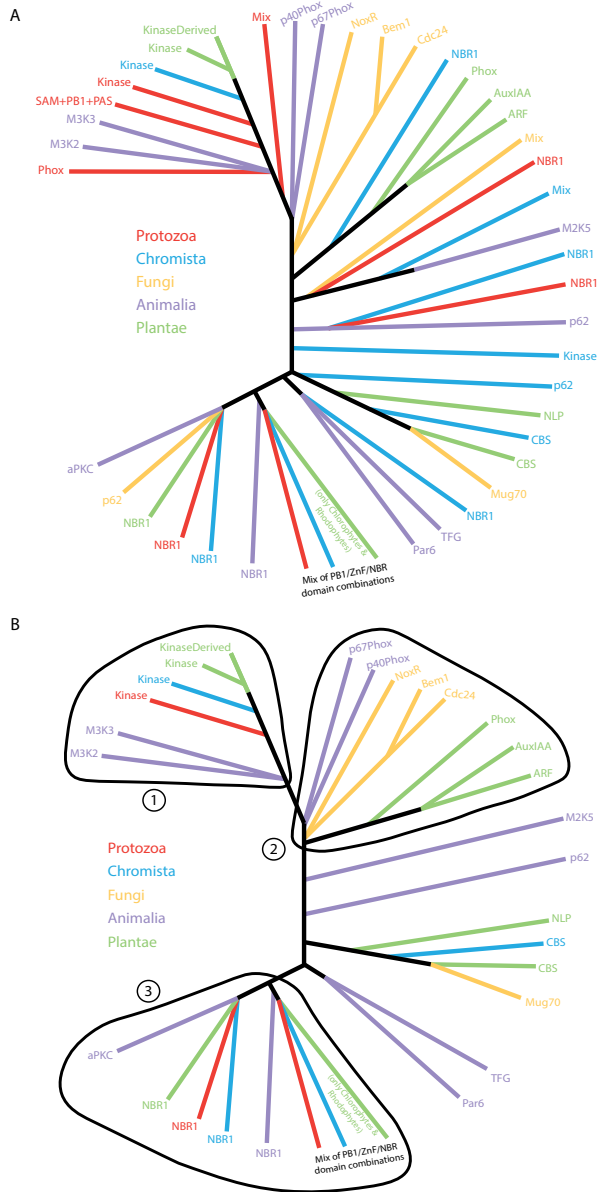


Figure S2: Complete (A) and simplified (B) illustration of the unrooted tree with the PB1 domains from all five kingdoms. Orthologs from each kingdom are represented with each colour as indicated: Protozoa in 'red', Chromista in 'blue', Fungi in 'orange', Animalia in 'purple' and Plantae in 'green'. The groups outlined with continuous lines indicated with numbers 1, 2 and 3 represent the probable ancestral copies in LECA corresponding to Kinase, Phox and NBR1 groups respectively. 'Mix' indicates a combination of (partial) PB1 domains with other domains in random. Full version of the tree with taxa names and domain information can be found at iTOL: <https://itol.embl.de/shared/dolfweijers>.

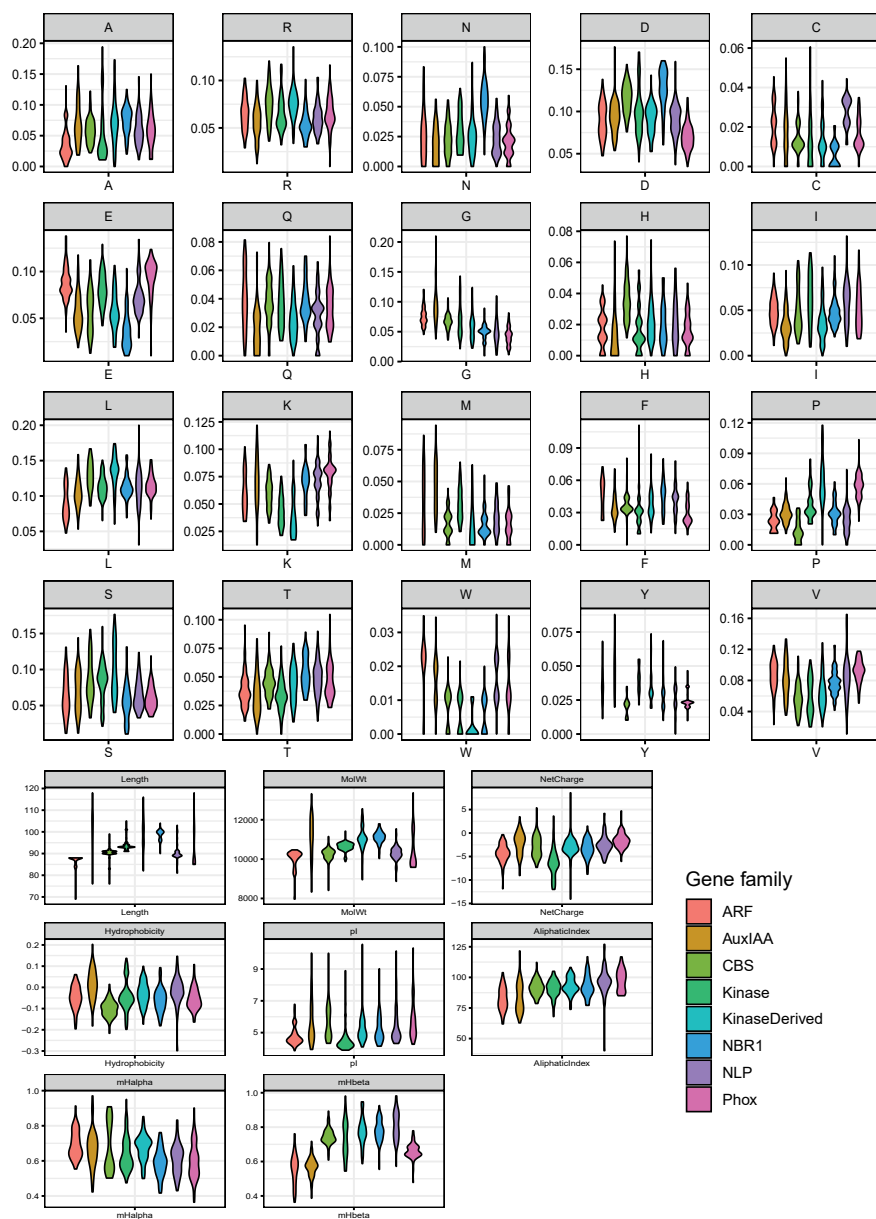


Figure S3: Violin plots showing the descriptive stats of the 28 descriptors/variables used Random Forest (RF) classification. Colours in each plot represent one gene family as shown in the legend.

Supplementary Tables

Table S1: List of domains (present in Fig. 1) with their short name and full name and InterproID.

ShortName	FullName (InterProID)
PB1	PB1 domain (IPR000270)
Kinase	Protein kinase domain (IPR000719)
WD40	WD40 repeat (IPR001680)
TPR	Tetratricopeptide repeat (IPR019734)
ZnF	Zinc finger, ZZ-type (IPR000433)
CentralDomain	Next to BRCA1, central domain (IPR032350)
UBA	Ubiquitin-associated domain (IPR015940)
CBS	CBS domain (IPR000644)
RWPRK	RWP-RK domain (IPR003035)
DH	Dbl homology (DH) domain (IPR000219)
PH	Pleckstrin homology domain (IPR001849)
SH3	SH3 domain (IPR001452)
PX	Phox homologous domain (IPR001683)
kDAG	Phorbol ester/diacylglycerol-binding domain (IPR002219)
PDZ	PDZ domain (IPR001478)
ARF-DBD/B3	B3 DNA binding domain (IPR003340)
ARF-DBD/ARF	Auxin response factor (IPR010525)
Aux/IAA-I	Domain-I or EAR motif
Aux/IAA-II	Domain-II or DEGRON motif

Table S2: List of identifiers of the PB1 domain containing proteins from four species of land plants (Marchantia, Physcomitrella, Amborella and Arabidopsis) used for the sequence alignment and logo construction.

GeneFamily	Marchantia	Physcomitrella	Amborella	Arabidopsis
ARF	Mapoly0019s0045.1	Pp3c1_14480V3.1	evm_27.model.AmTr_v1.0_scaffold00007.382	AT1G59750
	Mapoly0011s0167.1	Pp3c1_40270V3.1	evm_27.model.AmTr_v1.0_scaffold00016.128	AT5G62000
	Mapoly0075s0050.1	Pp3c2_25890V3.1	evm_27.model.AmTr_v1.0_scaffold00021.210	AT5G60450
		Pp3c4_12970V3.3	evm_27.model.AmTr_v1.0_scaffold00025.251	AT1G19850
		Pp3c4_13010V3.3	evm_27.model.AmTr_v1.0_scaffold00029.187	AT1G30330
		Pp3c5_9420V3.1	evm_27.model.AmTr_v1.0_scaffold00057.126	AT5G20730
		Pp3c6_21370V3.1	evm_27.model.AmTr_v1.0_scaffold00092.36	AT5G37020
		Pp3c13_4720V3.1	evm_27.model.AmTr_v1.0_scaffold00148.24	AT4G23980
		Pp3c14_16990V3.10	evm_27.model.AmTr_v1.0_scaffold00155.56	AT2G28350
		Pp3c16_6100V3.1	evm_27.model.AmTr_v1.0_scaffold00211.4	AT2G46530
		Pp3c17_19900V3.1		AT1G34310
		Pp3c27_60V3.1		AT1G34170
		Pp3c9_21330V3.1		AT1G35540

ARF		Pp3c15_9710V3.1		AT1G35520
				AT4G30080
				AT3G61830
				AT1G19220
				AT1G35240
				AT1G34410
				AT1G34390
AuxIAA	Mapoly0013s0010.1	Pp3c24_6610V3.1	evm_27.model.AmTr_v1.0_scaffold00002.512	AT4G14560
	Mapoly0034s0017.1	Pp3c8_14720V3.1	evm_27.model.AmTr_v1.0_scaffold00002.514	AT3G23030
			evm_27.model.AmTr_v1.0_scaffold00019.282	AT1G04240
			evm_27.model.AmTr_v1.0_scaffold00039.160	AT5G43700
			evm_27.model.AmTr_v1.0_scaffold00045.141	AT1G15580
			evm_27.model.AmTr_v1.0_scaffold00056.118	AT1G52830
			evm_27.model.AmTr_v1.0_scaffold00109.120	AT3G23050
			evm_27.model.AmTr_v1.0_scaffold00122.5	AT2G22670
			evm_27.model.AmTr_v1.0_scaffold00184.12	AT5G65670
				AT1G04100
				AT4G28640
				AT1G04550
				AT2G33310
				AT4G14550
				AT3G04730
				AT1G04250
				AT1G51950
				AT3G15540
				AT2G46990
				AT3G16500
				AT4G29080
				AT5G25890
				AT4G32280
				AT3G62100
				AT3G17600
				AT2G01200
				AT5G57420
				AT1G15050
CBS	Mapoly0179s0023.1	Pp3c1_14290V3.1	evm_27.model.AmTr_v1.0_scaffold00013.41	AT5G63490
		Pp3c1_14310V3.1	evm_27.model.AmTr_v1.0_scaffold00017.72	AT2G36500
		Pp3c11_15160V3.1		AT3G52950
		Pp3c2_26110V3.1		AT5G50640
		Pp3c7_10070V3.1		
Kinase	Mapoly0013s0150.1	Pp3c15_24250V3.1	evm_27.model.AmTr_v1.0_scaffold00004.293	AT1G04700
		Pp3c9_25280V3.1	evm_27.model.AmTr_v1.0_scaffold00019.236	AT1G16270
			evm_27.model.AmTr_v1.0_scaffold00026.90	AT1G79570
			evm_27.model.AmTr_v1.0_scaffold00039.196	AT2G35050
			evm_27.model.AmTr_v1.0_scaffold00081.26	AT3G24715





Kinase				AT3G46920
				AT5G57610
Kinase-derived			evm_27.model.AmTr_v1.0_scaffold00007.258	AT1G25300
			evm_27.model.AmTr_v1.0_scaffold00046.178	AT1G70640
			evm_27.model.AmTr_v1.0_scaffold00049.221	AT2G01190
			evm_27.model.AmTr_v1.0_scaffold00109.93	AT3G18230
				AT3G26510
				AT3G48240
				AT4G05150
				AT5G09620
				AT5G16220
				AT5G49920
				AT5G63130
				AT5G64430
NBR1	Mapoly0100s0042.1	Pp3c11_16970V3.1	evm_27.model.AmTr_v1.0_scaffold00049.238	AT4G24690
		Pp3c7_8990V3.1		
NLP	Mapoly0083s0040.1	Pp3c12_2070V3.1	evm_27.model.AmTr_v1.0_scaffold00058.115	AT2G17150
		Pp3c15_9180V3.1	evm_27.model.AmTr_v1.0_scaffold00066.150	AT4G35270
		Pp3c17_4370V3.1	evm_27.model.AmTr_v1.0_scaffold00080.66	AT4G38340
		Pp3c17_4375V3.1		AT1G20640
		Pp3c19_2670V3.1		AT1G76350
		Pp3c19_2720V3.1		AT1G64530
		Pp3c22_6360V3.1		AT4G24020
		Pp3c22_6370V3.1		AT2G43500
		Pp3c9_14600V3.1		AT3G59580

Table S3: Confusion matrix from the Random Forest (RF) model. Diagonal values represent the correctly classified PB1's, and others represent the mis-classified category. The column 'class.error' represents the classification error for that particular class of PB1 domains as shown in Fig. 5A.

	ARF	AuxIAA	CBS	Kinase	KinaseDerived	NBR1	NLP	Phox	class.error
ARF	99	0	1	0	0	0	0	0	0.01
AuxIAA	4	95	1	0	0	0	0	0	0.05
CBS	2	1	92	1	0	2	2	0	0.08
Kinase	0	1	0	95	4	0	0	0	0.05
KinaseDerived	1	0	2	7	87	0	1	2	0.13
NBR1	0	2	2	1	1	72	0	0	0.07
NLP	0	0	3	0	1	0	95	1	0.05
Phox	0	2	0	0	0	0	2	96	0.04

References

- Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., et al. (2012). The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* 59, 429–493.
- Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van De Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.* 46, D1190–D1196.
- Bienz, M. (2014). Signalosome assembly by domains undergoing dynamic head-to-tail polymerization. *Trends Biochem. Sci.* 39, 487–495.
- Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32.
- Burke, R.M., and Berk, B.C. (2015). The Role of PB1 Domain Proteins in Endothelial Cell Dysfunction and Disease. *Antioxid. Redox Signal.* 22, 1243–1256.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421.
- Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973.
- Carpenter, E.J., Matasci, N., Ayyampalayam, S., Wu, S., Sun, J., Yu, J., Jimenez Vieira, F.R., Bowler, C., Dorrell, R.G., Gitzendanner, M.A., et al. (2019). Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). *Gigascience* 8, 1–7.
- van Dop, M., Fiedler, M., Mutte, S., de Keijzer, J., Olijslager, L., Albrecht, C., Liao, C.-Y., Janson, M.E., Bienz, M., and Weijers, D. (2020). A conserved biochemical paradigm underlies cell polarity across multicellular kingdoms. *Cell*. in press.
- Fujiwara, M., Uemura, T., Ebine, K., Nishimori, Y., Ueda, T., Nakano, A., Sato, M.H., and Fukao, Y. (2014). Interactomics of Qa-SNARE in *Arabidopsis thaliana*. *Plant Cell Physiol.* 55, 781–789.
- Guan, P., Ripoll, J.-J., Wang, R., Vuong, L., Bailey-Steinitz, L.J., Ye, D., and Crawford, N.M. (2017). Interacting TCP and NLP transcription factors control plant responses to nitrate availability. *Proc. Natl. Acad. Sci.* 114, 2419–2424.
- Guilfoyle, T.J., and Hagen, G. (2012). Getting a grasp on domain III/IV responsible for Auxin Response Factor-IAA protein interactions. *Plant Sci.* 190, 82–88.
- Hagen, G., and Guilfoyle, T. (2002). Auxin-responsive gene expression: genes, promoters and regulatory factors. *Plant Mol. Biol.* 49, 373–385.
- Hirano, Y., Yoshinaga, S., Takeya, R., Suzuki, N.N., Horiuchi, M., Kohjima, M., Sumimoto, H., and Inagaki, F. (2005). Structure of a Cell Polarity Regulator, a Complex between Atypical PKC and Par6 PB1 Domains. *J. Biol. Chem.* 280, 9653–9661.
- Honbou, K., Minakami, R., Yuzawa, S., Takeya, R., Suzuki, N.N., Kamakura, S., Sumimoto, H., and Inagaki, F. (2007). Full-length p40phox structure suggests a basis for regulation mechanism of its membrane binding. *EMBO J.* 26, 1176–1186.
- Ito, T., Matsui, Y., Ago, T., Ota, K., and Sumimoto, H. (2001). Novel modular domain PB1 recognizes pc motif to mediate functional protein-protein interactions. *EMBO J.* 20, 3938–3946.
- Joberty, G., Petersen, C., Gao, L., and Macara, I.G. (2000). The cell-polarity protein Par6 links Par3 and atypical protein kinase C to Cdc42. *Nat. Cell Biol.* 2, 531–539.
- Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240.
- Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., and Jermini, L.S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589.
- Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
- Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMET-SP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* 12, e1001889.
- Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858.
- Kirkin, V., McEwan, D.G., Novak, I., and Dikic, I. (2009). A Role for Ubiquitin in Selective Autophagy. *Mol. Cell* 34, 259–269.
- Konishi, M., and Yanagisawa, S. (2019). The role of protein-protein interactions mediated by the PB1 domain of NLP transcription factors in nitrate-inducible gene expression. *BMC Plant Biol.* 19, 90.
- Korasick, D.A., Westfall, C.S., Lee, S.G., Nanao, M.H., Dumas, R., Hagen, G., Guilfoyle, T.J., Jez, J.M., and Strader,

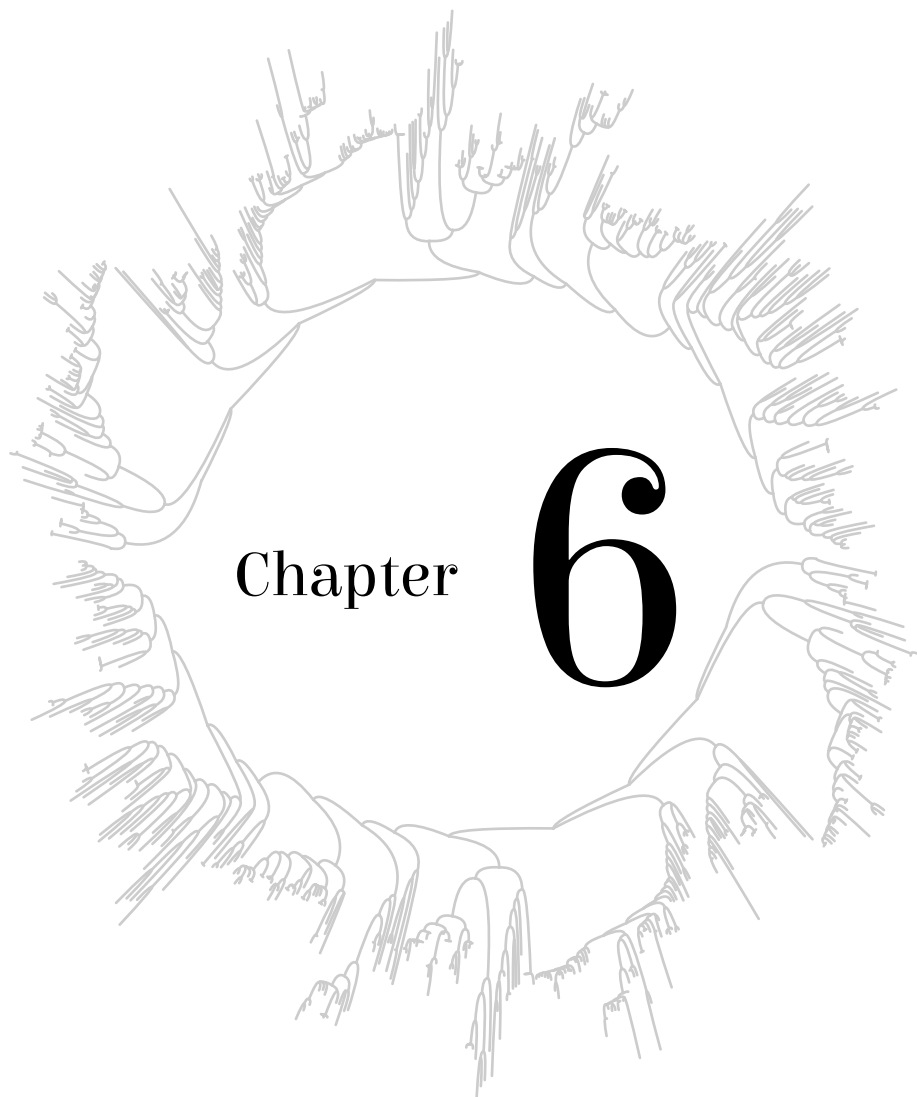




- L.C. (2014). Molecular basis for AUXIN RESPONSE FACTOR protein interaction and the control of auxin response repression. *Proc. Natl. Acad. Sci.* 111, 5427–5432.
- Kurth, E.G., Peremyslov, V. V., Turner, H.L., Makarova, K.S., Iranzo, J., Mekhedov, S.L., Koonin, E. V., and Dolja, V. V. (2017). Myosin-driven transport network in plants. *Proc. Natl. Acad. Sci. U. S. A.* 114, E1385–E1394.
- Lamark, T., Perander, M., Outzen, H., Kristiansen, K., Øvervatn, A., Michaelsen, E., Bjørkøy, G., and Johansen, T. (2003). Interaction Codes within the Family of Mammalian Phox and Bem1p Domain-containing Proteins. *J. Biol. Chem.* 278, 34568–34581.
- Lambeth, J.D. (2004). NOX enzymes and the biology of reactive oxygen. *Nat. Rev. Immunol.* 4, 181–189.
- Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2, 18–22.
- Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., et al. (2014). Data access for the 1,000 Plants (1KP) project. *Gigascience* 3, 17.
- Moscat, J., Diazmeo, M., And Wooten, M. (2007). Signal integration and diversification through the p62 scaffold protein. *Trends Biochem. Sci.* 32, 95–100.
- Mu, X., and Luo, J. (2019). Evolutionary analyses of NIN-like proteins in plants and their roles in nitrate signaling. *Cell. Mol. Life Sci.* 76, 3753–3764.
- Müller, S., Kursula, I., Zou, P., and Wilmanns, M. (2006). Crystal structure of the PB1 domain of NBR1. *FEBS Lett.* 580, 341–344.
- Mutte, S.K., Kato, H., Schiefel, C., Melkonian, M., Wong, G.K.-S., and Weijers, D. (2018). Origin and evolution of the nuclear auxin response system. *Elife* 7, e33399.
- Nakamura, K., and Johnson, G.L. (2003). PB1 Domains of MEKK2 and MEKK3 Interact with the MEK5 PB1 Domain for Activation of the ERK5 Pathway. *J. Biol. Chem.* 278, 36989–36992.
- Nakamura, R., Sumimoto, H., Mizuki, K., Hata, K., Ago, T., Kitajima, S., Takeshige, K., Sakaki, Y., and Ito, T. (1998). The PC motif: A novel and evolutionarily conserved sequence involved in interaction between p40(phox) and p67(phox), SH3 domain-containing cytosolic factors of the phagocyte NADPH oxidase. *Eur. J. Biochem.* 251, 583–589.
- Nanao, M.H., Vinos-Poyo, T., Brunoud, G., Thévenon, E., Mazzoleni, M., Mast, D., Lainé, S., Wang, S., Hagen, G., Li, H., et al. (2014). Structural basis for oligomerization of auxin transcriptional regulators. *Nat. Commun.* 5, 3617.
- Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.
- Noda, Y., Kohjima, M., Izaki, T., Ota, K., Yoshinaga, S., Inagaki, F., Ito, T., and Sumimoto, H. (2003). Molecular Recognition in Dimerization between PB1 Domains. *J. Biol. Chem.* 278, 43516–43524.
- Osorio, D., Rondón-Villareal, P., and Torres, R. (2015). Peptides: A package for data mining of antimicrobial peptides. *R J.* 7, 4–14.
- Piya, S., Shrestha, S.K., Binder, B., Stewart, C.N., and Hewezi, T. (2014). Protein-protein interaction and gene co-expression maps of ARFs and Aux/IAAs in Arabidopsis. *Front. Plant Sci.* 5, 1–9.
- Roccatò, E., Pagliardini, S., Cleris, L., Canevari, S., Formelli, F., Pierotti, M.A., and Greco, A. (2003). Role of TFG sequences outside the coiled-coil domain in TRK-T3 oncogenic activation. *Oncogene* 22, 807–818.
- Ruggiero, M.A., Gordon, D.P., Orrell, T.M., Bailly, N., Bourgoin, T., Brusca, R.C., Cavalier-Smith, T., Guiry, M.D., and Kirk, P.M. (2015). A higher level classification of all living organisms. *PLoS One* 10, 1–60.
- Sievers, F., and Higgins, D.G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* 27, 135–145.
- Sumimoto, H. (2008). Structure, regulation and evolution of Nox-family NADPH oxidases that produce reactive oxygen species. *FEBS J.* 275, 3249–3277.
- Sumimoto, H., Kamakura, S., and Ito, T. (2007). Structure and function of the PB1 domain, a protein interaction module conserved in animals, fungi, amoebas, and plants. *Sci. STKE* 2007, 1–16.
- Suzuki, A. (2006). The PAR-aPKC system: lessons in polarity. *J. Cell Sci.* 119, 979–987.
- Svenning, S., Lamark, T., Krause, K., and Johansen, T. (2011). Plant NBR1 is a selective autophagy substrate and a functional hybrid of the mammalian autophagic adapters NBR1 and p62/SQSTM1. *Autophagy* 7, 993–1010.
- Takemoto, D., Kamakura, S., Saikia, S., Becker, Y., Wrenn, R., Tanaka, A., Sumimoto, H., and Scott, B. (2011). Polarity proteins Bem1 and Cdc24 are components of the filamentous fungal NADPH oxidase complex. *Proc. Natl. Acad. Sci.* 108, 2861–2866.
- Terasawa, H. (2001). Structure and ligand recognition of the PB1 domain: a novel protein module binding to the PC motif. *EMBO J.* 20, 3947–3956.
- Trehin, C., Schrempp, S., Chauvet, A., Berne-Dedieu, A., Thierry, A.-M., Faure, J.-E., Negrutiu, I., and Morel, P. (2013). QUIRKY interacts with STRUBBELIG and PAL OF QUIRKY to regulate cell growth anisotropy during Arabidopsis gynoecium development. *Development* 140, 4807–4817.
- Weijers, D., and Wagner, D. (2016). Transcriptional Responses to the Auxin Hormone. *Annu. Rev. Plant Biol.* 67, 539–574.

- Xiao, N., Cao, D.S., Zhu, M.F., and Xu, Q.S. (2015). Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* 31, 1857–1859.
- Zientara-Rytter, K., and Sirko, A. (2014). Significant role of PB1 and UBA domains in multimerization of Joka2, a selective autophagy cargo receptor from tobacco. *Front. Plant Sci.* 5, 1–13.





Chapter 6

General discussion

The truth that does not need any proof is: ‘data is growing’. This is especially true in life sciences, with the advancement of technologies for next-generation and high-throughput analysis. Genomics is one such area that has seen and continues to see quantum leaps in data availability from sequencing genomes and transcriptomes of thousands of diverse species in the biosphere. With the ever-growing accuracy and speed of RNA-Seq, the domain of molecular phylogenetics is making advancements in the accuracy of inferences of relationships between genes and between species. Even though transcriptomes provide limited information about the gene content of an organism, it is largely sufficient to deduce the evolutionary relationships between homologs of a gene family (Delsuc et al., 2005; Wickett et al., 2014). Taking advantage of the relative ease of obtaining RNA-Seq data, over the last decade there has been a clear transition in area of phylogenetics from few reference-based genes to the large-scale multi-gene phylogenies, referred as phylogenomics.

The original algorithms in bioinformatic and evolutionary analysis were generated in times at which data was much more limited than at present. Thus, in parallel to the growing data, we need improved and robust analysis methods to deal with ever increasing datasets. There is particular need for new computational approaches when dealing with species that are spread over large evolutionary distances, or when the quality of data is limited, for example in large-scale sequencing projects. In this thesis, we have utilized two large-scale transcriptome sequencing datasets, One Thousand Plant Transcriptomes (OneKP; Carpenter et al., 2019) representing the kingdom Archaeplastida and the Marine Micro Eukaryote Transcriptome Sequencing Project (MMETSP; Keeling et al., 2014), majorly representing the kingdom Chromista.

In Chapter 2, we developed a bioinformatic pipeline to build phylogenetic trees utilizing the above-mentioned datasets, and to reconstruct the ancestral states of various gene families involved in synthesis, metabolism and signal transduction of the phytohormone auxin. This hormone is found in a wide range of unicellular as well as multicellular organisms, including bacteria (Amin et al., 2015). Despite its ubiquitous presence, no genetic basis for its role in growth and development was found in any other kingdoms except Archaeplastida. In contrast, a clear genetic basis and major routes of auxin synthesis, transport, metabolism and signal transduction are relatively well established in plants. Hence, it is important to understand why, when and how these systems evolved.

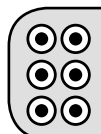
Previous studies have attempted to understand the origin and evolution of various auxin pathways (Finet et al., 2013; De Smet et al., 2011; Viaene et al., 2013). However, the majority of these studies relied on small numbers of species and gene families to make evolutionary inferences. Nevertheless, it was suggested that the majority of the orthologs of land plant auxin pathways were found in charophytes, but absent in chlorophytes. Indeed, recently published charophyte genomes confirmed the presence of orthologs of various pathways that were earlier thought to be specific to land plants (Jiao et al., 2019; Nishiyama et al., 2018). In Chapter 3, we also found that the single homologs of auxin biosynthetic gene families (TAA and YUC) were found in charophytes but not in the chlorophytes. A similar trend was observed for the

GH3 auxin metabolism genes. Conversely, the auxin binding ABP1 protein was found even in chlorophytes, and we found that all residues within its auxin binding pocket are deeply conserved. This indicates that the ABP1 might bind to auxin and mediate auxin-dependent responses in the ancestor of Viridiplantae.

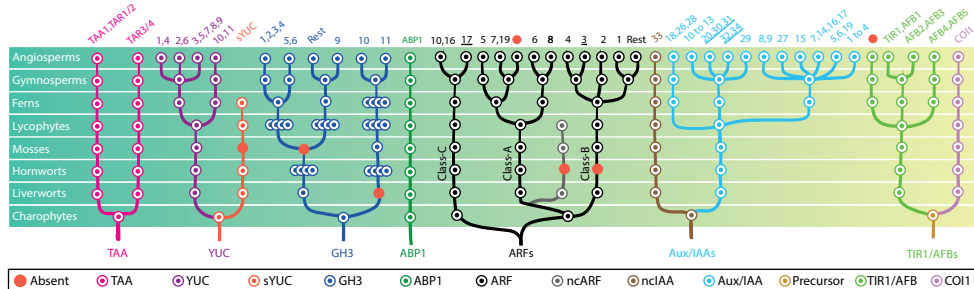
First evidence of (polar) auxin transport in algae was observed in *Chara corallina*, of the class Charophyceae (Boot et al., 2012). Recently it was shown that *Klebsormidium flaccidum*, a species that belongs to the Klebsormidiophyceae that diverged earlier than Charophyceae, also has a functional PIN protein that can mediate polar auxin transport when expressed in land plants (Skokan et al., 2019). However, the role of PIN proteins in growth and development as well as in polar auxin transport of these algal species is yet to be elucidated. Another family of proteins homologous to PIN proteins, PIN-LIKES (PILS), were identified even in chlorophyte algae (Barbez et al., 2012; Feraru et al., 2012). PILS are involved in intracellular auxin transport and maintenance of homeostasis. Interestingly, PILS are localized to the endoplasmic reticulum (ER), where the ABP1 protein involved in non-genomic responses is also majorly localized. Taken together, these results suggest that the origin of ER localized transport facilitators and non-genomic responses predates the genomic nuclear auxin responses as well as synthesis and polar auxin transport.

After finding out the origin or first appearance of homologous genes of auxin pathways in algal species, we studied the divergence in land plants. The only gene families that never diverged in land plants were the TAA and ABP1 families and nIAA proteins (Fig. 1). Among the land plants, the first step towards increased complexity is found in the vascular plants, where all other gene families studied diverged, followed by further diversification in seed plants. However, independent losses of orthologs in some clades or species were observed in many gene families, except TAA, ABP1 and Aux/IAA families (Fig. 1). Given that auxin mutants show abnormal phenotypes, with major defects in developmental programs, there is a possibility that species or clades with specific gene losses might have a reduced capacity of various auxin-dependent processes. Nevertheless, the effect of these independent losses (for e.g. nARF or class-B ARF in hornworts) for the development and physiology of those species is yet to be elucidated through genetic experiments.

Apart from the above-mentioned ‘independent losses’, we also identified ‘abnormal duplications’ in multiple instances. For example, gymnosperms have three ancestral copies of Aux/IAA proteins whereas in angiosperms, they are more than tripled in number. This finding can be attributed to the increased regulation and interaction capacities with ARF proteins in contribution to potential tissue-specific functions in angiosperms. On the other hand, strange duplication patterns were identified in GH3 proteins especially in hornworts and mosses. It is certain that more GH3 proteins could lead to increased capacity for homeostasis, but given that there is no increase in auxin synthesis gene families (Fig. 1), it is unusual to see these variations specific to homeostasis. It is however possible that the duplicated GH3 copies are expressed in specific tissues, and thus do not directly increase the cellular capacity to inactivate auxin. We



also identified ‘orphan’ lineages, that were discontinued from genomes at a certain point in evolution, as seen in sYUC and ncARF clades (Fig. 1). These orphan lineages appear to be lost after attaining inflated complex systems in the ‘main’ lineages of vascular and seed plants.



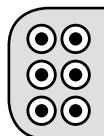
A key finding of this study is the identification of “non-canonical” components. These could be divided into two categories: non-canonical by origin and non-canonical by necessity. Sub-classes or orthologs that are “non-canonical by origin” are the ncARF and ncIAA proteins. Even though ncARF is derived from a class-A ARFs, and lacks the DNA binding domain, it is not present in all land plants (Fig. 1). However, genetic analysis showed that the ncARF has a positive contribution to auxin-dependent growth in *Marchantia* (Chapter 4). ncIAA lacks both the EAR motif and degron sequence that are needed for recruitment of the TPL co-repressor and for protein degradation through ubiquitin pathway, respectively. Conversely, ncIAA is present in all the land plant phyla, but our study in the early diverged liverwort *Marchantia* did not reveal a clear function in auxin responses (Chapter 4; Mutte et al., 2018). In *Arabidopsis*, recently it was shown that the ncIAA ortholog (IAA33), stabilized by MITOGEN-ACTIVATED PROTEIN KINASE 14 (MPK14), negatively regulates auxin signaling by interacting with ARF10 and ARF16 and by competing with the canonical IAA5 (Lv et al., 2019). Hence, it is evident that plants evolved multiple ways of regulation by invoking these non-canonical components in auxin response.

Supporting this hypothesis, we have identified some Kinases, that also contain a PB1 domain and are conserved across multiple-kingdoms (Chapter 5). It is worth investigating if the auxin-dependent PB1 domain-based kinase phosphorylation could induce rapid auxin responses. Recently it was shown that auxin can inhibit root growth within 2min in *Arabidopsis*, which is probably too fast to involve a transcriptional response (Fendrych et al., 2018). A possibility is that it requires the above-mentioned kinase-PB1 proteins.

Apart from the Aux/IAA family, the ARF family has also generated proteins that are 'non-canonical by necessity': ARF3 (ETTIN) and ARF17. In contrast to ncARF, these two proteins contain the DNA-binding domain but lack the C-terminal PB1 domain. Interestingly, these proteins evolved their atypical nature only in angiosperms, from a canonical ARF protein in the common ancestor with gymnosperms. ARF3/ETT is important for proper gynoecium development and controls various other processes by interacting with other proteins (Kelley et al., 2012; Sessions and Zambryski, 1995). Strikingly, recent NMR analysis showed that ARF3 can directly bind auxin to modulate gene expression by changing the chromatin states to promote gynoecium development (Kuhn et al., 2019). This is the first evidence to show that an ARF transcription factor can be a hormone receptor, which opens new avenues to investigate other receptors of auxin that mediate growth and development in plants.

On the other hand, ARF17 is important for anther dehiscence and pollen wall pattern formation in angiosperms (Xu et al., 2019; Yang et al., 2013). ARF17, and other class-C ARFs, are known to be regulated by microRNA miR160 both in *Arabidopsis* and *Marchantia* (Flores-Sandoval et al., 2018; Mallory et al., 2005). Moreover, as the class-C proteins evolved independently from class-A/B in charophytes (Chapter 4), they appear to also have an independent regulation based on miR160 instead of the canonical auxin-TIR1-Aux/IAA pathway (Flores-Sandoval et al., 2018). Further, the only class-C ARF in *Marchantia* (MpARF3) also regulates different target genes than class-A MpARF1 (Kato et al., 2019). Interestingly, the same study also showed that class-A ARF activates downstream auxin-responsive genes, which is antagonized by an auxin-independent class-B ARF that represses common target genes (Kato et al., 2019). All these recent evidences indicate multiple novel ways of auxin-TIR1 independent genetic regulation, that each control plant growth and development.

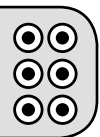
The components of the NAP have ancestral homologs in charophytes, but all the components are likely functional in early-diverged land plants. Various feedback loops are involved in auxin response pathways in angiosperms. Upon auxin treatment, the transcription of Aux/IAA, PIN and GH3 genes are all upregulated, whereas YUC is downregulated. By the activation of PIN efflux carriers, excess auxin is pumped out of the cell (Vieten et al., 2005). By increasing the levels of GH3 proteins in the cell, there is a decrease in levels of active IAA form, as it is converted to conjugated forms by GH3 proteins (Ludwig-Müller, 2011). Auxin biosynthesis enzyme YUC is transcriptionally repressed whereby the high cellular auxin levels stop internal auxin biosynthesis (Suzuki et al., 2015). In a similar way, Aux/IAA genes are transcriptionally upregulated upon auxin treatment by ARF proteins through direct interaction with Aux/IAA



gene promoters (Krogan and Berleth, 2015). But the feedback loops that are known to be active in angiosperms appear to be evolved later in vascular plants (Chapter 4). Comparative transcriptome analysis of various charophytes, bryophytes and a fern, upon auxin treatment, showed that not all ‘feedback’ families are regulated in the same way in either charophytes or bryophytes. In ferns however, there is a clear activation of Aux/IAA and GH3 transcripts and repression of YUC transcripts (Chapter 4), suggesting a robust feedback mechanism involving all these gene families did not exist prior to the emergence of vascular plants. It also suggests that bryophytes might have different feedback mechanism. Nevertheless, they appear to have a common module to activate transcription. We have identified C2HDZ and WIP genes to be commonly upregulated in all land plants tested, which is consistent even in flowering plants (Mutte et al., 2018). This indicates that along with the evolution of ARF proteins in land plants, the same auxin response cis-element (AuxRE; TGTC core) based gene activation might be ‘fixed’ in all land plants. Further promoter analysis of auxin response genes in the early diverged species should confirm these findings.

Hormone perception through F-box mediated receptors is not only specific to auxin, but also include other hormones such as jasmonate, gibberellic acid and strigolactones. In our study we found that the auxin (TIR1) and jasmonate (COI1) have dedicated receptors, only in land plants but not in green algae. Correspondingly, the clade (GH3.11/JAR1) specific to the synthesis of active form of jasmonate (Ja-Ile), also emerged in land plants (Fig. 1; Chapter 3). This indicates that ligand and receptor have parallel evolution in land plants. Interestingly, liverworts lack the ortholog for this clade, indicating that COI1 may not bind to JA-Ile, as it is not produced due to the lack of ortholog for GH3.11/JAR1. Indeed, it has been confirmed recently that *Marchantia* COI1 binds to the precursor of JA-Ile, and the complete synthesis and signaling pathways for jasmonate are functional in vascular plants (Monte et al., 2018; Pratiwi et al., 2017). Similar to jasmonate, canonical pathways for gibberellin and strigolactones synthesis and signal transduction are also observed in vascular, seed and flowering plants but not in bryophytes, due to lack of some functional components in those pathways (Bythell-Douglas et al., 2017; Hernández-García et al., 2019; Walker et al., 2019). Taken together, these results indicate that auxin pathways are early diverged and functional in all land plants, suggesting auxin as the first and oldest complete canonical F-box dependent phytohormone pathway.

On the other hand, two-component histidine kinase signaling components of cytokinin and ethylene pathways evolved prior to land plants, representing system that is more ancient than the nuclear auxin pathway (Bowman et al., 2017; Jiao et al., 2019; Ju et al., 2015). In contrast, despite the presence of some ABA synthesis pathway components in charophytes, the ABA-dependent responses were found only in land plants and further enhanced in vascular plants by gaining additional receptors (Sun et al., 2019). Hence, it is evident that not only the complexity of existing pathways that might have shaped the evolution of novel morphological traits (vascular tissue, seeds and flowers) in land plants but also adaptation and innovation of novel components throughout plant evolution shaped these life-history traits. Even though it appears implausible to innovate so many gene families and mechanisms during a single transition from charophytes



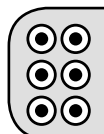
to land plants, this transition took more than 300 million years, which made these innovations possible. Hence finding the evolutionary intermediates, if extant, may reveal the order and timeline of events behind these innovations.

Acknowledgment

Sincere thanks to Shubhajit Das and Dr. Simon Lindhoud for proofreading the manuscript.

References

- Amin, S.A., Hmelo, L.R., van Tol, H.M., Durham, B.P., Carlson, L.T., Heal, K.R., Morales, R.L., Berthiaume, C.T., Parker, M.S., Djunaedi, B., et al. (2015). Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* 522, 98–101.
- Barbez, E., Kubeš, M., Rolčík, J., Béziat, C., Pěňčík, A., Wang, B., Rosquete, M.R., Zhu, J., Dobrev, P.I., Lee, Y., et al. (2012). A novel putative auxin carrier family regulates intracellular auxin homeostasis in plants. *Nature* 485, 119–122.
- Boot, K.J.M., Libbenga, K.R., Hille, S.C., Offringa, R., and van Duijn, B. (2012). Polar auxin transport: an early invention. *J. Exp. Bot.* 63, 4213–4218.
- Bowman, J.L., Kohchi, T., Yamato, K.T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., et al. (2017). Insights into Land Plant Evolution Garnered from the *Marchantia polymorpha* Genome. *Cell* 171, 287–304.
- Bythell-Douglas, R., Rothfels, C.J., Stevenson, D.W.D., Graham, S.W., Wong, G.K.S., Nelson, D.C., and Bennett, T. (2017). Evolution of strigolactone receptors by gradual neo-functionalization of KAI2 paralogues. *BMC Biol.* 15, 1–21.
- Cao, M., Chen, R., Li, P., Yu, Y., Zheng, R., Ge, D., Zheng, W., Wang, X., Gu, Y., Gelová, Z., et al. (2019). TMK1-mediated auxin signalling regulates differential growth of the apical hook. *Nature* 568, 240–243.
- Carpenter, E.J., Matasci, N., Ayyampalayam, S., Wu, S., Sun, J., Yu, J., Jimenez Vieira, F.R., Bowler, C., Dorrell, R.G., Gitzendanner, M.A., et al. (2019). Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). *Gigascience* 8, 1–7.
- Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375.
- Fendrych, M., Akhmanova, M., Merrin, J., Glanc, M., Hagihara, S., Takahashi, K., Uchida, N., Torii, K.U., and Friml, J. (2018). Rapid and reversible root growth inhibition by TIR1 auxin signalling. *Nat. Plants* 4, 453–459.
- Feraru, E., Vosolobé, S., Feraru, M., Petrášek, J., and Kleine-Vehn, J. (2012). Evolution and structural diversification of PILS putative auxin carriers in plants. *Front. Plant Sci.* 3, 227.
- Finet, C., Berne-Dedieu, A., Scutt, C.P., and Marlétaz, F. (2013). Evolution of the ARF gene family in land plants: Old domains, new tricks. *Mol. Biol. Evol.* 30, 45–56.
- Flores-Sandoval, E., Eklund, D.M., Hong, S.F., Alvarez, J.P., Fisher, T.J., Lampugnani, E.R., Golz, J.F., Vázquez-Lobo, A., Dierschke, T., Lin, S.S., et al. (2018). Class C ARFs evolved before the origin of land plants and antagonize differentiation and developmental transitions in *Marchantia polymorpha*. *New Phytol.* 218, 1612–1630.
- Hernández-García, J., Briones-Moreno, A., Dumas, R., and Blázquez, M.A. (2019). Origin of Gibberellin-Dependent Transcriptional Regulation by Molecular Exploitation of a Transactivation Domain in DELLA Proteins. *Mol. Biol. Evol.* 36, 908–918.
- Jiao, C., Sørensen, I., Sun, X., Sun, H., Behar, H., Alseekh, S., Philippe, G., Lopez, K.P., Sun, L., Reed, R., et al. (2019). The Genome of the Charophyte Alga *Penium margaritaceum* Bears Footprints of the Evolutionary Origins of Land Plants. *BioRxiv* 835561.
- Ju, C., Van De Poel, B., Cooper, E.D., Thierer, J.H., Gibbons, T.R., Delwiche, C.F., and Chang, C. (2015). Conservation of ethylene as a plant hormone over 450 million years of evolution. *Nat. Plants* 1, 1–7.
- Kato, H., Mutte, S.K., Suzuki, H., Crespo, I., Das, S., Radoeva, T., Fontana, M., Yoshitake, Y., Hainiwa, E., Berg, W. van den, et al. (2019). Design principles of a minimal auxin response system. *BioRxiv* 760876.
- Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMET-SP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* 12, e1001889.



- Kelley, D.R., Arreola, A., Gallagher, T.L., and Gasser, C.S. (2012). ETTIN (ARF3) physically interacts with KANADI proteins to form a functional complex essential for integument development and polarity determination in Arabidopsis. *Development* 139, 1105–1109.
- Krogan, N.T., and Berleth, T. (2015). The identification and characterization of specific ARF-Aux/IAA regulatory modules in plant growth and development. *Plant Signal. Behav.* 10, e992748–e992748.
- Kuhn, A., Harbrough, S.R., McLaughlin, H.M., Kepinski, S., and Østergaard, L. (2019). Direct ETTIN-auxin interaction controls chromatin state in gynoecium development. *BioRxiv* 863134.
- Ludwig-Müller, J. (2011). Auxin conjugates: their role for plant development and in the evolution of land plants. *J. Exp. Bot.* 62, 1757–1773.
- Lv, B., Yu, Q., Liu, J., Wen, X., Yan, Z., Hu, K., Li, H., Kong, X., Li, C., Tian, H., et al. (2019). Non-canonical AUX/IAA protein IAA33 competes with canonical AUX/IAA repressor IAA5 to negatively regulate auxin signaling. *EMBO J.* e101515.
- Mallory, A.C., Bartel, D.P., and Bartel, B. (2005). MicroRNA-Directed Regulation of Arabidopsis AUXIN RESPONSE FACTOR17 Is Essential for Proper Development and Modulates Expression of Early Auxin Response Genes. *Plant Cell* 17, 1360–1375.
- Monte, I., Ishida, S., Zamarreño, A.M., Hamberg, M., Franco-Zorrilla, J.M., García-Casado, G., Gouhier-Darimont, C., Reymond, P., Takahashi, K., García-Mina, J.M., et al. (2018). Ligand-receptor co-evolution shaped the jasmonate pathway in land plants. *Nat. Chem. Biol.* 14, 480–488.
- Mutte, S.K., Kato, H., Rothfels, C., Melkonian, M., Wong, G.K.-S., and Weijers, D. (2018). Origin and evolution of the nuclear auxin response system. *Elife* 7, e33399.
- Nishiyama, T., Sakayama, H., de Vries, J., Buschmann, H., Saint-Marcoux, D., Ullrich, K.K., Haas, F.B., Vanderstraeten, L., Becker, D., Lang, D., et al. (2018). The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. *Cell* 174, 448–464.
- Pratiwi, P., Tanaka, G., Takahashi, T., Xie, X., Yoneyama, K., Matsuura, H., and Takahashi, K. (2017). Identification of Jasmonic Acid and Jasmonoyl-Isoleucine, and Characterization of AOS, AOC, OPR and JAR1 in the Model Lycophyte *Selaginella moellendorffii*. *Plant Cell Physiol.* 58, 789–801.
- Sato, A., and Yamamoto, K.T. (2008). Overexpression of the non-canonical Aux/IAA genes causes auxin-related aberrant phenotypes in Arabidopsis. *Physiol. Plant.* 133, 397–405.
- Sessions, R.A., and Zambryski, P.C. (1995). Arabidopsis gynoecium structure in the wild and in ettin mutants. *Development* 121, 1519–1532.
- Skokan, R., Medvecká, E., Viaene, T., Vosolsobě, S., Zwiewka, M., Müller, K., Skůpa, P., Karady, M., Zhang, Y., Janacek, D.P., et al. (2019). PIN-driven auxin transport emerged early in streptophyte evolution. *Nat. Plants* 5, 1114–1119.
- De Smet, I., Voß, U., Lau, S., Wilson, M., Shao, N., Timme, R.E., Swarup, R., Kerr, I., Hodgman, C., Bock, R., et al. (2011). Unraveling the Evolution of Auxin Signaling. *Plant Physiol.* 155, 209–221.
- Sun, Y., Harpazi, B., Wijerathna-Yapa, A., Merilo, E., de Vries, J., Michaeli, D., Gal, M., Cuming, A.C., Kollist, H., and Mosquna, A. (2019). A ligand-independent origin of abscisic acid perception. *Proc. Natl. Acad. Sci.* 116, 24892–24899.
- Suzuki, M., Yamazaki, C., Mitsui, M., Kakei, Y., Mitani, Y., Nakamura, A., Ishii, T., Soeno, K., and Shimada, Y. (2015). Transcriptional feedback regulation of YUCCA genes in response to auxin levels in Arabidopsis. *Plant Cell Rep.* 34, 1343–1352.
- Viaene, T., Delwiche, C.F., Rensing, S.A., and Friml, J. (2013). Origin and evolution of PIN auxin transporters in the green lineage. *Trends Plant Sci.* 18, 5–10.
- Vieten, A., Vanneste, S., Wiśniewska, J., Benková, E., Benjamins, R., Beeckman, T., Luschnig, C., and Friml, J. (2005). Functional redundancy of PIN proteins is accompanied by auxin-dependent cross-regulation of PIN expression. *Development* 132, 4521–4531.
- Walker, C.H., Siu-Ting, K., Taylor, A., O'Connell, M.J., and Bennett, T. (2019). Strigolactone synthesis is ancestral in land plants, but canonical strigolactone signalling is a flowering plant innovation. *BMC Biol.* 17, 70.
- Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. U. S. A.* 111, E4859–E4868.
- Xu, X.-F., Wang, B., Feng, Y.-F., Xue, J.-S., Qian, X.-X., Liu, S.-Q., Zhou, J., Yu, Y.-H., Yang, N.-Y., Xu, P., et al. (2019). AUXIN RESPONSE FACTOR17 Directly Regulates MYB108 for Anther Dehiscence. *Plant Physiol.* 181, 645–655.
- Yang, J., Tian, L., Sun, M.-X., Huang, X.-Y., Zhu, J., Guan, Y.-F., Jia, Q.-S., and Yang, Z.-N. (2013). AUXIN RESPONSE FACTOR17 Is Essential for Pollen Wall Pattern Formation in Arabidopsis. *Plant Physiol.* 162, 720–731.

Summary

Auxin is a key phytohormone for growth and development across many plant species. With the advances in next-generation sequencing, and ever-growing data of genomes and transcriptomes in the last decade across all plant lineages, helped us infer the detailed origin and evolution of plant auxin biology in this thesis. **Chapter 1** provides an overview of the major forms of life on earth, from an evolutionary perspective. We also provided detailed insight about multiple phyla in the plant evolution, and the corresponding morpho-physiological innovations. We then introduced various proteins and pathways involved in the synthesis, transport, homeostasis and signal transduction of auxin, followed by the scope of this thesis.

Large ‘omics’ data resources, especially transcriptomics as a result of advancement in RNA-Seq technologies, generated more than 1300 transcriptomes from more than a thousand species over a billion years of plant evolution. Even though the transcriptome data is inherently limited than the genome data, it is possible to estimate ancestral copies i.e. the minimal gene complement in the ancestor of that particular lineage under consideration. While many methods were developed that either focus on combining single copy genes to estimate species relationships, or focus on a specific gene family, needs substantial modification to apply as a general method to other gene families. Hence, in **Chapter 2** we have developed a simple yet effective methodology to reconstruct the origin and evolution of various genes families across all plant lineages, including the algal ancestors.

We first tested this protocol in **Chapter 3** to study the evolution of auxin biosynthesis gene families Tryptophan aminotransferase of Arabidopsis (TAA) and YUCCA (YUC). This has not only confirmed the previous findings, that TAA and its homologous Alliinase proteins had a common ancestor in charophyte algae, but also revealed that the core TAA clade has never duplicated in their ancestral states during the plant evolution. YUC family in early diverged species contain a sister clade that is lost in the recently diverged species, angiosperms. Further, we exploited the evolution of Gretchen Hagen 3 (GH3) protein family involved in the auxin homeostasis that conjugate amino acids to hormones. This analysis revealed a striking correlation between the emergence of auxin or jasmonate specific GH3 family members to the evolution of specialized hormone receptors at the onset of land plant evolution. Moreover, we also showed that this method could be applied to proteins with unknown or novel domains, by studying the evolution and annotation of SOSEKI proteins. Auxin elicit both genomic and non-genomic responses, through nuclear auxin pathway and Auxin binding protein 1 (ABP1), respectively. Hence, we also studied the origin and evolution of ABP1, that appear to have evolved in red algae, with auxin binding capacity predicted to appear as early as in chlorophytes.

In **Chapter 4**, we focused on the nuclear auxin pathway, the major auxin signal transduction pathway in plants. We studied both the origin and deep evolution of Auxin Response Factor (ARF), Auxin/Indole-3-acetic acid (Aux/IAA) and Transport inhibitor response 1/Auxin-signaling F-box (TIR1/AFB) gene families. Based on the evolutionary patterns, we hypothesized that genomic response pathway components were evolved in charophytes but functional only in

land plants. We tested this hypothesis by studying the auxin response capacity of green algae, hornworts, liverworts, mosses and ferns, comparing the auxin treated transcriptome to the untreated plants using RNA-Seq. Indeed, we found that the auxin response genes were conserved across land plants, but not in green algae, confirming the predictions from evolutionary analysis. To our surprise, we also found deeply conserved non-canonical components (ncARF and ncIAA), which we tested for their role in auxin dependent responses by studying the mutants in early diverged bryophyte, *Marchantia*. Interestingly, ncARF appeared as a positive contributor to the auxin dependent growth and development.

A key step in the nuclear auxin pathway, is the interaction between ARF and Aux/IAA proteins through C-terminal Phox and Bem1 (PB1) domain. During our search for homologs of these two gene families, we found many other proteins in plants that are neither ARFs nor Aux/IAAs, and contain a PB1 domain. As PB1 domain is also identified in protozoans, fungi as well as in animals, in **Chapter 5** we studied if the PB1 domain is originated in Last Eukaryote Common Ancestor (LECA). Indeed, we found that LECA consisted preferably three PB1 copies that diverged further and gave rise to multiple copies in various kingdoms. We found co-evolved PB1 pairs that might interact, where we also found common conserved amino acids among these predicted interaction pairs, for e.g. Kinase and Kinase-derived PB1 domains. We have also successfully applied amino acid descriptor based Random Forest classification to differentiate various PB1 domains across land plants. Taken together, these results revealed the broader presence of PB1 domain containing proteins in plants, with insights into their deep evolution.

Finally, in **Chapter 6** we discussed how this study contributes to what is known about the evolution of various auxin pathway components and their functions. The results in this thesis collectively show the step-wise origins of auxin biology in green algae and subsequent functionality in land plants. We also highlighted questions that still remain and synchronize these findings with the on-going efforts to understand 'non-canonical' routes of auxin action in plants.

Acknowledgements

As Darwin said, Evolution is ‘descent with modification’. Evolution is everywhere, in plants, animals, microbes, proteins, cities, technology... What not, everything evolves. Some changes happen at a scale that we can quantify, some are almost invisible, some are negligible while some are obvious! In my case, all these things happened too! My professional skills have evolved and continue to do so, to a great extent that I never thought, I came to a country that I never thought would travel to and I am living a life that I never dreamt of! All these happened and are happening because of help and support from all the people that I came across in my life. Every single person that I saw, met, or heard of, have an impact on my life.

I always believed that a good mentor is extremely crucial, both personally and professionally. I am extremely lucky to have such a bunch of mentors. I have reached this pinnacle of my student career, successfully, due to two such special people – Dolf and Ajay.

Dolf, you are such an amazing scientist, an awesome person and a great mentor. You make sure that your students are doing good both personally and professionally. I am superbly impressed by your managerial qualities. You gathered a nice team of people, who are great to work with. The atmosphere that you created with well-chosen and nicely crafted people creates nice harmony in the work place. Thank you for making me a part of this beautiful experience, and I am glad that you gave me another opportunity to extend this experience for few more years. Thank you for all your inputs and help. I would not have reached this point without your guidance.

Ajay, first of all I would like to thank you for being a great friend than a mentor. During my final year Bachelors’ degree, when you visited our lab, talked to me for hardly two minutes and immediately gave me an offer to come to Philippines. I am still surprised, how was that possible! You have always been very supportive and the first person to help me in any situation. I can never forget the amazing dinners, parties and best moments that I spent with your family. I am always thankful for your continuous support.

I have also learned a lot not only from my mentors, but also being a supervisor myself to Lisa, Rubaiat and Samuel. I tried my best to be a good mentor. I enjoyed the mentoring process and learned a lot from you all.

I would like to thank my paranymphs, Shubhajit and Kuan-Ju. Shubhajit, you are a quiet officemate and I always enjoy talking some random stuff to break that silence. Kuan-Ju, you always have time to help others and your perseverance is at a whole different level, the best part I like most. Thank you for your help and inspiration.

I would like to extend sincere thanks to my officemates: Cathy and Willy, and all other members of Plant Development team: Simon, Juan, Keita, Heidi, Dasha, Juriaan, Ping and Mattia. I am really looking forward to work with all the new members of the group - Joao, Vera, Victor, Andriy, Polet and Mariska. Mark, thank you for all the discussions and importantly a big thanks for helping me buy my first car. My special thanks to Prasad, for all the spiritual, scientific,

religious and non-sense conversations in 'telugu'. Sometimes emotions and information are best conveyed when you talk in your mother tongue. I am lucky that I could share everything without losing intensity.

I would like to thank all other members of the Biochemistry: Laura, Sacco, Iraes, Jan-Willem, Daan, Bel, Pilar, Adrie, Elwira, Sergio, Sjeff, Carlo and Jacques. It is always nice to have great colleagues around.

I can never forget the help of previous lab members and association with them. A very big thanks to 'quiet and intelligent' Hiro. What a collaboration we had! We have worked together on many aspects of 'Marchantia' and 'evolution'. I have learned a lot from you Hiro. Thank you for all your support and knowledge! I would also like to acknowledge the help of our amazing alumni: Tanya, Dong, Joakim, Ale, Shunsuke, Bert, Jos, Sebastien, Liao, Maritza, Nicole, Margo and Thijs. It was great working with you all. Willem, thanks for believing in my work and giving me a chance to work with Matthias. Though briefly, I really enjoyed the part of applying my knowledge to fungal genomes from an industrial biotech perspective! Pleasure working with both you and Matthias. Thank you Tao and Eric for nice discussions about synteny. I strongly believe that one day we will merge the worlds of both deep and broad evolution.

A 'thanks' will not be sufficient enough to thank my friends: Charan, Mahesh, Suresh and Reddy. You guys are just amazing! I never felt that I am far away from home because you guys always make sure to keep me happy. Charan and Mahesh, we are friends for last 20 years and I am sure this is going to be a lifelong friendship.

Last but not the least, well, when I like a dish very much - I always try to keep a part of it separately, to eat at the end. Because that 'best' part stays longer with me. The best part of my life is my family. Caring mom and supporting dad – Manju and Kondalu – I dedicate not only this book but also my success to you both. You both taught me the most important lessons of life - be kind, be helpful and be strong. Akka and Bava – Madhavi and Sateesh – can't thank you enough for taking care of me always. I never missed home because you were always there for me. Saanvi and Lasya, I love you both a lot. I am sure I am going to have a tough time in future, because I may never understand your English accent. Spoorthi, I am lucky to have you as my better half. It's been an amazing year of togetherness and understanding. I am sure it is going to be the same for the rest of life. Your support in the final crucial stages of PhD made my journey even smoother. A heartfelt special thanks to my extended family - Sravya, athayya and mamayya. Srinu babai, you always made sure that I would have the latest computational power needed for my job. Thank you for all your support babai! I always admire the affection you have for us. Finally, I would like to thank all other family members in 'Mutte', 'Vangaru' and 'Chepuru' clades.

Curriculum Vitae

Sumanth Kumar Mutte was born on the 16th of August 1990 in Kothapatnam, Andhra Pradesh, India. In 2005 he completed his secondary education at Sun Shine Public School and in 2007, higher secondary education at Sri Chaitanya Junior College in Ongole. Later, he moved to Coimbatore, India to pursue B.Tech degree in Bioinformatics at Tamil Nadu Agricultural University (TNAU). During his studies he became interested in various bioinformatics tools to analyze genomic data and *in silico* structural analysis of proteins. Particularly he was very much favored in utilizing public data resources to understand biological insights through large-scale omics data analysis. After obtaining a bachelors' degree in Bioinformatics from TNAU, in 2011, he moved to Philippines to apply his knowledge to rice research at International Rice Research Institute (IRRI), a CGIAR institution based in Los Banos. There he worked on identifying candidate genes in Quantitative Trait Loci (QTL) for grain yield under drought stress. In 2013, he moved to Wageningen, the Netherlands to pursue M.Sc in Bioinformatics. For his major thesis, he worked on understanding the transcriptome regulation during heat stress recovery in *C. elegans* at the Laboratory of Nematology, Wageningen University under the guidance of Dr. Basten Snoek in the group of Dr. Jan Kammenga. For his minor thesis, he joined Prof. Dolf Weijers lab at the Laboratory of Biochemistry in Wageningen. There he worked on understanding the evolution of Auxin Response Factors and gained novel insights into the deep evolution of these proteins in plants. During this project he was fascinated by the evolution of plants. Hence, after obtaining the Masters' degree in Bioinformatics, he decided to continue working on the same project as a Bioinformatician and later as a PhD student with Prof. Weijers. During this trajectory, he worked on various small projects, along with his main project to understand the evolution of auxin response pathway components in plants, resulting in this thesis. Sumanth wants to continue as a bioinformatician and stay connected to the field of evolution by continuing as a post-doctoral researcher at the Laboratory of Biochemistry focusing on omics data analysis and integration.



Publications

Mutte, S. K., & Weijers, D. (2020). Deep Evolutionary History of the Phox and Bem1 (PB1) Domain Across Eukaryotes. *Scientific Reports*, 10, 3797. DOI: 10.1038/s41598-020-60733-9.

Mutte, S. K., & Weijers, D. (2020). High-resolution and deep phylogenetic reconstruction of ancestral states from large transcriptomic data sets. *Bio-Protocol*, in press.

Kato, H., **Mutte, S.K.**, Suzuki, H., Crespo, I., Das, S., Radoeva, T., Fontana, M., Yoshitake, Y., Hainiwa, E., van den Berg, W., Lindhoud, S., Hohlbein, J., Borst, J.W., Boer, R., Nishihama, R., Kohchi, T., Ishizaki, K., Weijers, D. (2020) Design principles of a minimal auxin response system. *Nature Plants*, in press.

Li, F-W., Nishiyama, T., Waller, M., Frangedakis, E., Keller, J., Li, Z., Fernandez-Pozo, N., Barker, M., Bennett, T., Blazquez, M., Cheng, S., Cuming, A., de Vries, J., de Vries, S., Delaux, P-M., Issa Diop, S., Harrison, C., Hauser, D., Hernández-García, J., Kirbis, A., Meeks, J., Monte, I., **Mutte, S.K.**, Neubauer, A., Quandt, D., Robison, T., Shimamura, M., Rensing, S., Villarreal J., Weijers, D., Wicke, S., Wong, G., Sakakibara, K., Szoenyi, P. (2020) Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts. *Nature Plants*. DOI: 10.1038/s41477-020-0618-2.

van Dop, M.*, Fiedler, M.*, **Mutte, S.**, de Keijzer, J., Olijslager, L., Albrecht, C., Liao, C-Y., Janson, M., Bienz, M., Weijers, D. (2020). DIX Domain Polymerization Drives Assembly of Plant Cell Polarity Complexes. *Cell*, 180(3), 427–439. DOI: 10.1016/j.cell.2020.01.011.

Lu, K.-J.*, van 't Wout Hofland, N.*, Mor, E., **Mutte, S.**, Abrahams, P., Kato, H., Vandepoele, K., Weijers, D., De Rybel, B. (2020). Evolution of vascular plants through redeployment of ancient developmental regulators. *Proceedings of the National Academy of Sciences*, 117(1), 733–740. DOI: 10.1073/pnas.1912470117.

Mutte, S. K.*, Kato, H.*, Rothfels, C., Melkonian, M., Wong, G. K.-S., Weijers, D. (2018). Origin and evolution of the nuclear auxin response system. *eLife*, 7, e33399. DOI: 10.7554/eLife.33399.

Frommhagen, M., **Mutte, S. K.**, Westphal, A. H., Koetsier, M. J., Hinz, S. W. A., Visser, J., Vincken, J-P., Weijers, D., van Berkel, W. J. H., Gruppen, H., Kabel, M. A. (2017). Boosting LPMO-driven lignocellulose degradation by polyphenol oxidase-activated lignin building blocks. *Biotechnology for Biofuels*, 10, 121. DOI: 10.1186/s13068-017-0810-4.

Palovaara, J.*, Saiga, S.*, Wendrich, J. R., van 't Wout Hofland, N., van Schayck, J. P., Hater, F., **Mutte, S.**, Sjollem, J., Boekschoten, M., Hooiveld, G., Weijers, D. (2017). Transcriptome dynamics revealed by a gene expression atlas of the early Arabidopsis embryo. *Nature Plants*, 3(11), 894–904. DOI: 10.1038/s41477-017-0035-3.

Dixit, S., Kumar Biswal, A., Min, A., Henry, A., Oane, R. H., Raorane, M. L., Longkumer, T., Pabuayon, I. M., **Mutte, S. K.**, Vardarajan, A. R., Miro, B., Govindan, G., Albano-Enriquez, B., Pueffeld, M., Sreenivasulu, N., Slamet-Loedin, I., Sundarvelandian, K., Tsai, Y-C., Raghuvanshi, S., Hsing, Y-I. C., Kumar, A., Kohli, A. (2015). Action of multiple intra-QTL genes concerted around a co-localized transcription factor underpins a large effect QTL. *Scientific Reports*, 5, 15183. DOI: 10.1038/srep15183.

Raorane, M. L., Pabuayon, I. M., Vardarajan, A. R., **Mutte, S. K.**, Kumar, A., Treumann, A., Kohli, A. (2015). Proteomic insights into the role of the large-effect QTL qDTY12.1 for rice yield under drought. *Molecular Breeding*, 35(6), 139. DOI: 10.1007/s11032-015-0321-6.

Raorane, M. L., **Mutte, S. K.**, Vardarajan, A. R., Pabuayon, I. M., Kohli, A. (2013). Protein SUMOylation and plant abiotic stress signaling: in silico case study of rice RLKs, heat-shock and Ca²⁺-binding proteins. *Plant Cell Reports*, 32(7), 1053–1065. DOI: 10.1007/s00299-013-1452-z.

* These authors contributed equally

Education Statement of the Graduate School

Experimental Plant Sciences



Issued to: **Sumanth Kumar Mutte**
 Date: **15 April 2020**
 Group: **Biochemistry**
 University: **Wageningen University & Research**

1) Start-Up Phase	<u>date</u>	<u>cp</u>
► First presentation of your project Origin and evolution of Auxin Response Factors	13 May 2016	1.5
► Writing or rewriting a project proposal Computational approaches to understand the nuclear auxin response system	07 Nov 2017	1.5
► MSc courses		

Subtotal Start-Up Phase

3.0

2) Scientific Exposure	<u>date</u>	<u>cp</u>
► EPS PhD student days Gel2Gether, Soest, NL	15-16 Feb 2018	0.6
Gel2Gether, Soest, NL	11-12 Feb 2019	0.6
► EPS theme symposia Theme 1 - Developmental Biology of Plants, Wageningen, NL	21 Jan 2016	0.3
Theme 1 - Developmental Biology of Plants, Wageningen, NL	30 Jan 2018	0.3
Theme 1 - Developmental Biology of Plants, Leiden, NL	31 Jan 2019	0.3
Theme 4 - Genome Biology, Wageningen, NL	13 Dec 2019	0.3
► Lunteren Days and other national platforms Annual Meeting Experimental Plant Sciences, Lunteren, NL	11-12 Apr 2016	0.6
Annual Meeting Experimental Plant Sciences, Lunteren, NL	10-11 Apr 2017	0.6
Annual Meeting Experimental Plant Sciences, Lunteren, NL	09-10 Apr 2018	0.6
Annual Meeting Experimental Plant Sciences, Lunteren, NL	08-09 Apr 2019	0.6
BioSB - Dutch Bioinformatics and Systems Biology Conference, Lunteren, NL	15-16 May 2018	0.6
► Seminars (series), workshops and symposia Joint meeting: groups of Prof. Dolf Weijers and Prof. Ben Scheres, Wageningen	22 Jan 2016	0.2
Seminar: Dr. Mark Estelle	06 Apr 2016	0.1
Seminar: Dr. Jill Harrison	12 May 2016	0.1
Seminar: Dr. Helene Robert	03 Jun 2016	0.1
Seminar: Dr. Jose Gutierrez-Marcos	07 Mar 2016	0.1
Seminar: Dr. Chun-Ming Liu	21 Jun 2016	0.1
Seminar: Dr. Sabine Muller	09 Sep 2016	0.1
Seminar: Dr. Lars Ostergaard	21 Dec 2016	0.1
Symposium: Molecules@WURK	8 Feb 2017	0.2
Seminar: Dr. Gerd Jurgens	11 May 2017	0.1
Seminar: Dr. Katharina Burstenbinder	01 Jun 2017	0.1
Seminar: Dr. Arnold Boersma	19 Jun 2017	0.1
Seminar: Dr. Manuel Juette	05 Sept 2017	0.1
Seminar: Next-Generation Sequencing, WUR	14 Sept 2017	0.1
Seminar: Prof. Venkatesan Sundaresan	30 Apr 2018	0.1
Seminar: Dr. Charles Delwiche	18 Jun 2018	0.1
Seminar: Dr. Victoria Mironova	27 Jun 2018	0.1
Seminar: Prof. Lucia Strader	10 Sep 2018	0.1
Symposium: A tribute to Life Sciences	25 Oct 2018	0.2
Symposium: 45 years of Yellow fever	02 Nov 2018	0.2
Symposium: The Brave New World of Smart Data & Semantics in the Life Sciences	24 Jan 2019	0.3
Molecular Life Sciences seminar: Prof. Ludwik Leibler	22 Nov 2018	0.1
Molecular Life Sciences seminar: Prof. Wolf Frommer	17 Jan 2019	0.1
Molecular Life Sciences seminar: Prof. Rudolf Zechner	21 Mar 2019	0.1
Seminar: Dr. Maheshi Dassanayake	20 May 2019	0.1
Seminar: Dr. Enrico Scarpella	19 Jun 2019	0.1
Seminar: Dr. Ari-Pekka Mähönen	19 Jun 2019	0.1
Symposium: Land plant evolution and Photosynthesis	20 Jun 2019	0.3
Symposium: Five years of Bioinformatics @ EPS	10 Jul 2019	0.2
Seminar: Dr. Minako Ueda	04 Sep 2019	0.1
Seminar: Prof. dr. Miguel Blazquez	08 Oct 2019	0.1
Seminar: Dr. Jill Harrison	08 Oct 2019	0.1
► Seminar plus		
► International symposia and congresses Auxin2016, Sanya, China	20-25 Oct 2016	1.6
AuxinWorkshop, Ghent, Belgium	28-30 Sep 2017	0.8
EMBO-meeting: New Shores in Land Plant Evolution, Lisbon, Portugal	20-23 Jun 2018	1.1
International Conference on Plant Growth Substances, Paris, France	25-29 Jun 2019	1.3
► Presentations Poster @ Auxin2016, Sanya, China	20-25 Oct 2016	1.0
Talk @ AuxinWorkshop, Ghent, Belgium	28-30 Sep 2017	1.0
Poster @ EMBO-meeting, Lisbon, Portugal	20-23 Jun 2018	1.0
Talk @ B-WISE seminar, Wageningen, NL	01 May 2018	1.0
Poster @ International Conference on Plant Growth Substances, Paris, France	25-29 Jun 2019	1.0
► 3rd year interview		
► Excursions		

Subtotal Scientific Exposure

19.3

CONTINUED ON NEXT PAGE

3) In-Depth Studies	<u>date</u>	<u>cp</u>
▶ Advanced scientific courses & workshops		
Introduction to Multiomics data integration, EBI, Cambridge, UK	20-23 Feb 2018	1.2
High-performance Computing (HPC) basic course, Wageningen, NL	17 May 2018	0.2
EPS PhD Summer school: Environmental Signaling in Plants, Utrecht, NL	26-28 Aug 2019	0.9
▶ Journal club		
Journal club - Plant Development group, Biochemistry, WUR	2016 - 2019	3.0
▶ Individual research training		

Subtotal In-Depth Studies

5.3

4) Personal Development	<u>date</u>	<u>cp</u>
▶ General skill training courses		
Research data management, Wageningen, NL	01-15 Mar 2018	0.5
Scientific writing, Wageningen, NL	13 Mar - 30 Apr 2018	1.8
EPS introduction course, Wageningen, NL	27 Mar 2018	0.3
Brain training, Wageningen, NL	05 Feb 2019	0.3
Last stretch of PhD programme, Wageningen, NL	14 May 2019	0.0
Career assessment, Wageningen, NL	13 Jun 2019	0.3
▶ Organisation of meetings, PhD courses or outreach activities		
▶ Membership of EPS PhD Council		

Subtotal Personal Development

3.2

5) Teaching & Supervision Duties	<u>date</u>	<u>cp</u>
▶ Courses		
Bioinformation technology (SSB20306)	2018, 2019	3.0
▶ Supervision of BSc/MSc students		
M.Sc thesis: Evolution of IPA pathway of auxin biosynthesis in plants; RN Akhand	Jul 2018	1.0
M.Sc thesis: Evolution of ROPs and ROPGEFs in plants; S Gebretsadkan	Sep 2018	1.0
M.Sc minor thesis: Effect of H3K27me3 on auxin responsive genes in Arabidopsis root; RN Akhand	Jan 2019	1.0

Subtotal Teaching & Supervision Duties

6.0

TOTAL NUMBER OF CREDIT POINTS*	36.8
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 presented in this thesis was performed at the Laboratory of Biochemistry, Wageningen University and Research, and was financially supported by a grant from the Netherlands Organization for Scientific Research (NWO; VICI 865.14.001).

Financial support from the Laboratory of Biochemistry for printing this thesis is gratefully acknowledged.

Cover design by Sumanth Kumar Mutte

Layout by Sumanth Kumar Mutte

Printed by ProefschriftMaken

