Phylogenomic analysis of the APETALA2 transcription factor subfamily across angiosperms reveals both deep conservation and lineage-specific patterns

Merijn H. L. Kerstens, M. Eric Schranz* and Klaas Bouwmeester

Biosystematics Group, Wageningen University and Research, Droevendaalsesteeg 1, Wageningen 6708 PB, the Netherlands

Received 6 January 2020; revised 30 April 2020; accepted 7 May 2020.

*For correspondence (e-mail eric.schranz@wur.nl).

SUMMARY

The APETALA2 (AP2) subfamily of transcription factors are key regulators of angiosperm root, shoot, flower and embryo development. The broad diversity of anatomical and morphological structures is potentially associated with the genomic dynamics of the AP2 subfamily. However, a comprehensive phylogenomic analysis of the AP2 subfamily across angiosperms is lacking. We combined phylogenetic and synteny analysis of distinct AP2 clades in the completed genomes of 107 angiosperm species. We identified major changes in copy number variation and genomic context within clades across lineages, and discuss how these changes may have contributed to the evolution of lineage-specific traits. Multiple AP2 clades show highly conserved patterns of copy number and synteny across angiosperms, while others are more dynamic and show distinct lineage-specific patterns. As examples of lineage-specific morphological divergence due to AP2 clade dynamics, we hypothesize that loss of PLETHORA1/2 in monocots correlates with the absence of taproots, whereas independent lineage-specific changes of PLETHORA4/BABY BOOM and WRINKLE1 genes in Brassicaceae and monocots point towards regulatory divergence of embryogenesis between these lineages. Additionally, copy number expansion of TOE1 and TOE3/AP2 in asterids is implicated with differential regulation of flower development. Moreover, we show that the genomic context of AP2s is in general highly specialized per angiosperm lineage. To our knowledge, this study is the first to shed light on the evolutionary divergence of the AP2 subfamily clades across major angiosperm lineages and emphasizes the need for lineage-specific characterization of developmental networks to understand trait variability further.

Keywords: synteny networks, phylogenomics, AP2/EREBP transcription factors, plant embryogenesis, root and shoot development, floral development.

INTRODUCTION

Angiosperms display an extraordinary diversity in anatomy and morphology across a large number of phylogenetic lineages, each with their own characteristic traits (Stevens, 2001; Leebens-Mack et al., 2019). The most distinctive lineages are the monocots and the eudicots, which can be characterized by having either one or two cotyledons respectively. Many crop and model species are distributed across the phylogenetic tree, such as in the rosids (i.e. soybean, strawberry, cassava, cucumber, Arabidopsis, cotton and orange), the asterids (tomato, coffee, carrot and lettuce) and in the monocots (wheat, corn, pineapple, banana, oil palm). Phenotypic diversity in angiosperm lineages can be largely explained by patterns of gene diversification. A major driver of genetic diversification in plants is gene duplication, which promotes functional redundancy (Van De Peer et al., 2009; Guo, 2013; Panchy et al., 2016). Various processes are thought to facilitate gene duplication, of which polyploidization events are regarded as having the most profound impact (Panchy et al., 2016). Polyploidization is often followed by the process of diploidization, which reshuffles genome structure by chromosomal rearrangements and massive gene loss with many duplicated genes returning to a single-copy state (Dodsworth et al., 2016). On a more local scale, numerous genes undergo a process of constant gene birth–death by tandem duplication, transposition and retroduplication (Panchy et al., 2016). Both global and local duplication mechanisms provide the raw material needed for gene diversification through neo- and subfunctionalization, thereby affecting angiosperm trait evolution (Ohno, 1970; Force et al., 1999; Lynch and Force, 2000).

© 2020 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
Retention of genomic position over the course of evolution, called synteny, is thought to be indicative of conservation of gene function (Dewey, 2011; Lv et al., 2011). Hence, studying the evolutionary history of gene synteny in related clades can reveal key events in the acquisition of novel traits caused by gene duplications and rearrangement of genomic context (Tang et al., 2008; Dewey, 2011; Jiao and Paterson, 2014). Recently, we developed a pipeline that exploits network clustering to study syntenic relationships between genes, overcoming challenges imposed by pairwise interspecies comparisons (Zhao and Schranz, 2017). With this approach, new evolutionary trends could be inferred for MADS-box and LEA gene families across 51 and 60 plant species respectively (Zhao et al., 2017; Artur et al., 2019). More recently, this approach was used to examine overall syntenic properties and genomic differences between angiosperms and mammals (Zhao and Schranz, 2019). These studies demonstrate the potential of network clustering to study synteny of diverse gene families simultaneously in multiple species.

The APETALA2/ETHYLENE-RESPONSIVE ELEMENT BINDING PROTEIN (AP2/EREBP) superfamily is one of the most prominent transcription factor families regulating plant development and stress responses (Riechmann and Meyerowitz, 1998). It can be divided into two main subfamilies: the AP2 subfamily, whose members contain two AP2 domains, and the EREBP subfamily containing a single AP2 domain (Riechmann and Meyerowitz, 1998; Shigyo et al., 2006). The AP2 subfamily can be further divided in the euANT, basALANT and euAP2 clades, of which members have been demonstrated in Arabidopsis thaliana to play key roles in diverse developmental processes (Kim et al., 2006). The euANT gene AINTEGUMENTA (ANT) was shown to regulate floral organ identity and shoot meristem maintenance (Elliott et al., 1996; Mudunukothge and Krizek, 2012). The six PLETHORA genes (PLT1, PLT2, PLT3/AIL6, PLT4/BABY BOOM, PLT5/AIL5 and PLT7/AIL7) are redundant regulators of multiple key processes, including root and shoot development, phyllotaxis/rhizotaxis and embryogenesis (Aida et al., 2004; Galinha et al., 2007; Smith and Long, 2010; Prasad et al., 2011; Hofhuis et al., 2013; Horstman et al., 2014). The basALANT genes WRI1, WRI3 and WRI4 (WRINKLED1/3/4) govern fatty acid metabolism in the embryo and flower (Carnac and Benning, 2004; To et al., 2012). Within the euAP2 clade, AP2 controls flower development and floral organ identity (Jofuku et al., 1994; Okamuro et al., 1997), while TOE1 and TOE3 (TARGET OF EA1/3) regulate flowering (Zhu and Helliwell, 2011; Zhang et al., 2015).

Having pivotal regulatory roles in multiple developmental processes, genome dynamics within the AP2 subfamily can have profound effects on the morphological characteristics of angiosperms. Thus far, the AP2 subfamily has only been studied in the context of a subset of subclades, a small number of species, or at low resolution in combination with the EREBP subfamily (Rashid et al., 2012; Song et al., 2013; Zumajo-Cardona and Pabón-Mora, 2016; Lakhwani et al., 2016; Najafi et al., 2018; Li et al., 2018; Leebens-Mack et al., 2019; Wang et al., 2019). To elucidate the evolution of the AP2 subfamily in greater detail, it is important to assess all subclades of this subfamily in an angiosperm-wide manner.

Here, we determined the dynamics in copy number variation and syntenic conservation of AP2 subfamily genes across all major angiosperm lineages. By combining phylogenetics and syntenic network clustering, we identified distinct genomic patterns in the AP2 subfamily, uncovering evolutionary forces that drive morphological and developmental diversity in angiosperms.

RESULTS AND DISCUSSION

Synteny conservation in AP2 subclades of angiosperms

Phylogenomic analysis of the AP2 subfamily was conducted by combining conventional phylogenetic and syntelogs methods (Figure 1b; Zhao and Schranz, 2019). The annotated proteins of 107 angiosperm species, including two basal eudicots, were mined for protein sequences containing at least two repeated AP2 domains using HMMER (Figure 1b). The resulting hits were termed ‘HMMERlogs’, homologous proteins that share identical domain compositions (Figure 1a). In total, we identified 2171 AP2 HMMERlogs across 107 angiosperm proteomes (Data S1). HMMERlog sequences were aligned, trimmed and then used for phylogenetic analysis. In parallel, we performed syntenic network analyses to identify AP2 genes localized in regions of similar genomic context, termed ‘syntelogs’ (Figure 1a). Of the 2171 AP2 HMMERlogs, 1570 are connected by synteny to at least one other HMMERlog (Data S2).

To gain insight into conservation and divergence of genomic context across the AP2 subfamily of angiosperms, syntenic connections were plotted on the phylogenetic tree (Figure 1 and Figure S1). Within the phylogenetic tree, three major clades and 14 distinct AP2 subclades can be distinguished (Figure 1c). Each AP2 subclade consists of highly interconnected HMMERlogs. This syntenic signal is in strong congruence with the observed phylogeny (Figure 1c), and supports the described gene phylogeny of the AP2 subfamily in angiosperms. An exception is the EREBPAP2-AP2 subclade consisting of HMMERlogs with two AP2 domains, but having overall higher homology with EREBP proteins. HMMERlogs of this subclade were not included in further analyses. One group of 17 AP2 HMMERlogs belongs to neither the euAP2 nor the basALANT clade, and was termed ‘ambiAP2’ for its ambiguity. BLASTP analysis showed that these ambiguous sequences resemble PLT1/2 HMMERlogs (Table S3).
Evolutionary divergence of APETALA2 transcription factors

Strong syntenic connections are found between HMMERlogs of four subclade pairs: ANT and PLT3/7; PLT1/2 andambiAP2; PLT5 and WRI3/4; and TOE3/AP2 and TOE1 (Figure 1c) respectively. Two of these subclade pairs contribute to the same plant developmental processes in Arabidopsis. ANT, PLT3 and PLT7 play key roles in shoot meristem maintenance (Mudunkothge and Krizek, 2012), whereas AP2, TOE1 and TOE3 regulate flowering (Jofuku et al., 1994; Okamuro et al., 1997; Zhu and Helliwell, 2011; Zhang et al., 2015). Similar patterns of shared synteny were detected for the Type II MADS-box subclade pairs AGL8-TM3, SEP1-SQUA and SEP3-FLC (Ruelens et al., 2013; Zhao et al., 2017). This suggests that the functions of these AP2 HMMERlogs are facilitated by a shared genomic context. The strong syntenic signal between the PLT1/2 andambiAP2 subclades is supported by the sequence similarity detected by BLASTP. As of yet, a potential shared function between the PLT5 and WRI3/4 subclades remains unclear.

AP2 copy number variation as potential driver of morphological diversity in angiosperms

We explored copy number variation in AP2 subclades to understand morphological divergence in angiosperms further (Figure 2 and Table S1). AP2 copy number and whole-genome duplication were found to be weakly correlated (Kendall’s $\tau \approx 0.287$), suggesting that the frequency of whole-genome duplication is not the major driving force underlying AP2 copy number variation observed in our dataset. Across most angiosperms, the PLT1/2 subclade displayed relatively little variation in copy number (i.e. one or two copies). However, all monocots lack HMMERlogs of this subclade, suggesting gene loss during early monocot evolution (Figure 2). As Arabidopsis PLT1 and PLT2 are crucial regulators of root development (Aida et al., 2004; Galinha et al., 2007), this process is potentially regulated in a different fashion in monocots. In support of this, the basal eudicot Amborella trichopoda contains a single PLT1/2 HMMERlog and has a taproot system (Trueba et al., 2016). Most eudicots have a taproot system; a primary root that develops directly from an embryonic root, out of which secondary roots emerge. In monocots, embryonic roots are aborted, inducing adventitious root formation that leads to the formation of a fibrous root system. While monocots lack PLT1/2 HMMERlogs, they do contain monocot-specific euANT-L HMMERlogs, which potentially assume the role of PLT1/2 in root development (Figure 2). Likewise, as Arabidopsis PLT4 also redundantly orchestrates root development (Galinha et al., 2007), the increase in the number of PLT4 HMMERlogs in monocots compared with eudicots (i.e. one to two versus two to three copies) could compensate for the loss of PLT1/2 genes. Regardless of their main root system, both eudicots and monocots form secondary roots. In Arabidopsis, PLT3, PLT5 and PLT7 govern rhizotaxis (Hoffhuis et al., 2013). For the PLT5 subclade, the number of copies is relatively constant across angiosperms, also between eudicots and monocots, although the latter lineage most often has two copies instead of one. The PLT3/7 subclade also shows relatively low copy number variation (i.e. one or two copies). The observed increase in copy number in several Brassicaceae is potentially a product of recent whole-genome duplications in this lineage. The relative conservation of these two

![Figure 1. Phylogenetic and syntenic relationships of the AP2 subfamily in angiosperms. (a) Schematic drawing illustrating the terms ‘HMMERlog’ (upper panel) and ‘syntelog’ (lower panel). HMMERlogs are defined as homologs with an identical domain composition (proteins A and B, both having two AP2 domains). Syntelogs are genes sharing conserved syntenic regions (e.g. genes A and B in regions 1 and 2). (b) Bioinformatics workflow for HMMERlog identification and phylogenomic analyses. Main analysis steps and used software tools are indicated. (c) Phylogenetic and synteny network analysis of AP2 HMMERlogs in 107 angiosperm genomes. Coloured lines indicate strong conservation of synteny between gene pairs within and between AP2 subclades. Weaker syntenic connections between subclades are shown in grey in Figure S1. Subclades are named following the Arabidopsis gene nomenclature. Subclades lacking Arabidopsis orthologues are annotated with the suffix -L (like). The phylogenetic tree was inferred by ML analysis using 1000 bootstraps.](image-url)
subclades might indicate that the secondary root developmental programme is conserved between eudicots and monocots, despite distinct differences in their main root system. Moreover, as PLT3, PLT5 and PLT7 also play key roles in Arabidopsis phyllotaxis (Prasad et al., 2011), our findings suggest that lateral organ positioning in general is conserved between these lineages.

Monocot and eudicot embryogenesis differ in the number of developed cotyledons, indicating that their embryo developmental programmes are probably regulated in a very different fashion. The question is thus raised, how developmental programmes are probably regulated in a very different fashion. The question is thus raised, how such differences are reflected by copy number variation in the AP2 subfamily. The absence of both PLT2 and PLT4 was shown to be embryo lethal in Arabidopsis (Galinha et al., 2007). As already pointed out, monocots lack PLT1/2 HMMERlogs, but have more PLT4 copies (Figure 2). This potentially buffers the loss of PLT1/2 function. Alternatively, monocot-specific euANT-L HMMERlogs may have defining roles during embryogenesis.

The WR1, WR13/4 and WR13/3/4-L subclades, consisting of transcriptional regulators of fatty acid synthesis (Cernac and Benning, 2004; To et al., 2012), are largely conserved in copy number across angiosperms, although not as much as PLT1/2, PLT3/7 and PLT5. However, Brassicaceae lack WRI1/3/4-L orthologues, suggesting an evolutionary loss of a distinct subclade for regulators of fatty acids.

Floral morphology in angiosperms is extremely diverse, as specified through the modular ‘ABC’ model. In comparison with other angiosperm lineages, asterids exhibit duplicate retention of a large number of TOE1 HMMERlogs and TOE3/AP2 HMMERlogs to a lesser extent. Potentially, this is linked to a divergent form of modular flower development in asterids. In support of this, major differences in regulation of floral organ patterning were found between Arabidopsis and Petunia hybrida, belonging to the rosid and asterid lineages respectively. In Arabidopsis, the transcription factor AP2 has an A-class function and antagonizes B- and C-class genes (Krokan et al., 2012). However, the closest homologues of AP2 in Petunia, that is, REPRESSOR OF B-FUNCTION1/2/3 (ROB1, ROB2 and ROB3), found in the TOE3/AP2 subclade, only antagonize B-class genes (Morel et al., 2017). The Petunia BLIND ENHANCER (BEN) gene has an identical role to Arabidopsis AP2 by repressing both B- and C-class genes, but instead belongs to the TOE1 subclade. Thus, whereas Arabidopsis TOE1 is a repressor of flowering time but does not regulate flower patterning, BEN acts in flower patterning but does not affect flowering time (Morel et al., 2017). In addition, BEN and ROB genes also regulate nectary size (Morel et al., 2018). This points out that expansion of the euAP2 clade in asterids has led to more complex patterning during flower development.

The number of ANT HMMERlogs is particularly variable across rosids. For example, ANT copy number is high in Fabaceae and Cucurbitaceae (3–4) and low in Brassicaceae (approximately 1). Although we are unable to pinpoint a specific trait that correlates with this, the defining role of ANT in cell proliferation in seeds, leaves and flowers (Krizek, 1999; Mizukami and Fischer, 2000; Confalonieri et al., 2014) suggests that ANT copy number might very well be involved in shaping the distinctive morphology of these structures in different angiosperm lineages.

AmbiAP2 HMMERlogs are present almost exclusively in the Rosaceae. Although they most resemble PLT1/2 HMMERlogs in terms of genomic context and local sequence similarity, it is unclear whether their role in root and/or embryo development is shared. Further research is necessary to dissect the role of the ambiaP2 subclade in Rosaceae development.

Angiosperm AP2 subfamily is impacted by ancestral and lineage-specific synteny

Extensive synteny conservation is often linked with preservation of gene function (Dewey, 2011; Lv et al., 2011). As such, we investigated synteny variation within AP2 subclades across angiosperms. For each angiosperm species,
we determined copy number variation within each syntelog cluster and performed hierarchical clustering (Figure 3a). In this way, 27 syntelog clusters were identified, including all subclades except ambioAP2, each representing a distinct genomic context (Figure 3b). Our data suggest that the genomic context of many AP2 subclades, despite their conserved roles in crucial developmental processes, varies across angiosperms. This demonstrates that synteny in the AP2 subfamily is largely lineage-specific, with only few subclades being deeply conserved across all angiosperms (Figure 3). The level of synteny conservation does not correlate with the number of HMMERlogs in a subclade. For example, the ANT subclade is the second largest in number of HMMERlogs, but only represented in a single syntelog cluster. This reassures that variation in syntelog cluster size is not merely a product of the number of HMMERlogs within a subclade.

As a next step, we quantified the level of synteny conservation across angiosperms (Figure 3c). Four AP2 subclades, each consisting of a single syntelog cluster, display extremely strong conservation of synteny across all angiosperm species. This includes the ANT, AIL1, PLT1/2 (although only present in eudicots) and PLT5 subclades, and proposes the existence of strict positional constraints to conserve gene function. As such, genes belonging to these subclades probably share the same core functions across all angiosperms. More dynamic patterns of synteny can be observed in the other AP2 subclades, each displaying the specific presence or absence of synteny in different angiosperm lineages. For instance, within the PLT3/7 subclade two separate genomic contexts can be observed, that is specific presence of synteny in monocots, and specific absence of synteny in eudicots. As PLT3 and PLT7 play key roles in phyllotaxis and rhizotaxis in Arabidopsis (Prasad et al., 2011; Hoffhuis et al., 2013), our data suggest that these processes are differently regulated in monocots. The PLT4 subclade exhibits markedly more lineage-specific synteny than its related subclades, including lineage specificity in Brassicaceae, monocots and Poales. This potentially influences the regulation of embryogenesis and root development in different angiosperm lineages. The WRI subclades are generally more dynamic concerning their genomic position in Brassicaceae, asterids and monocots, suggesting differential regulation of fatty acid synthesis in these lineages.

WRII and PLT4 are the only subclades that exhibit Brassicaceae-specific synteny. PLT4 transcriptionally regulates the LEC1-ABI3-FUS3-LEC2 (LAFL) transcription factor network that directs embryogenesis (Horstman et al., 2017). PLT4 and WRII are also direct downstream targets of individual components of this network (Baud et al., 2007; Mu et al., 2008; Wang and Perry, 2013; Pelletier et al., 2017). This finding highlights that Brassicaceae evolved unique mechanisms to regulate embryogenesis. This is reinforced by the fact that Arabidopsis LEC1, one of the four core LAFL genes, also belongs to a Brassicaceae-specific syntelog cluster (Figure S2). However, the other three LAFL genes do not exhibit Brassicaceae specificity (Figure S2). Future research will be needed to test the hypothesis that the unique syntenic positions of WRII, PLT4 and LEC1 affect the promoter and/or chromatin dynamics of these genes and thus create a regulatory network for Brassicaceae-specific embryo development.

Within the two TOE subclades, monocots and eudicots are in separate synteny clusters, implying regulation of distinct flower development in these major angiosperm lineages.
lineages. To our knowledge, these analyses are the first to define lineage-specific synteny relationships within the AP2 subfamily, and form a framework to unravel differential transcriptional regulation of various developmental programmes in angiosperms.

**Synteny of the AP2 subfamily is specialized per angiosperm lineage**

The AP2 subfamily displays a remarkably high level of lineage-specific synteny, particularly as this gene family orchestrates multiple developmental programmes in angiosperms. Therefore, we determined how this degree of lineage specificity relates to the presumably more dynamic sister subfamily of EREBP transcription factors, which play essential roles in regulating responses to biotic and abiotic stress (Licausi et al., 2013). In comparison with the AP2 subfamily, EREBP genes are more strongly interconnected by synteny, even though the ratio of syntelogs to HMMERlogs is identical (Figure 4a, left and middle panel), which is probably because of a larger number of EREBP syntelogs per cluster (Figure 4a, right panel). This implies that the EREBP subfamily has had more retention of duplicated genes or experienced duplication more often, but it does not reveal any information on the degree of lineage-specific synteny. Hence, we performed phylogenomic analyses of the EREBP subfamily and classified subclades as described by Nakano et al., 2006 (Figures S3–S6 and Table S2). Similar to the AP2 subfamily, EREBP subclades demonstrated copy number variation and lineage-specific synteny (Figures S4–S6). To quantify this, we calculated a cluster overlap index of all pairwise interspecies comparisons. These cluster overlap indices can be used to compare relative conservation of synteny between the AP2 and EREBP subfamilies (Figure 4b). Of both related families, synteny is less conserved across all angiosperms in the AP2 family. However, when the overall cluster overlap index is split into the five defined angiosperm lineages, conservation of AP2 subfamily synteny is stronger than conservation of EREBP subfamily synteny in Brassicaceae, monocots and Poales. This suggests that the AP2 genomic context is more specialized per lineage than that of the EREBPs.

**OUTLOOK**

Our study exemplifies that combining phylogenetics and synteny networking (Figure 1) is a powerful tool to investigate gene families, and the individual subclades therein, in large sets of species. Here, we have extended previous work by our group (Zhao et al., 2017; Zhao and Schranz, 2017; Artur et al., 2019; Zhao and Schranz, 2019) by separating the AP2 and EREBP subfamilies based on domain count, performing in-depth copy number and synteny analyses of individual subclades, and comparing the relative conservation of synteny between both subfamilies. We provide two lines of potential evidence that associate variations in the AP2 subfamily with angiosperm evolution. First, we reveal that angiosperm lineages differ in regard to gene copy number of phylogenetically distinct AP2 subclades (Figure 2). Secondly, we demonstrate that these subclades exhibit both deeply conserved and lineage-specific synteny (Figure 3). In addition, we show that the genomic contexts of AP2 genes are specialized per angiosperm lineage (Figure 4).

Through both a copy number and synteny perspective, we put forth a set of hypotheses that could be interesting to pursue in further research on the AP2 subfamily. In the PLT1/2 subclade, eudicots possess in general only one or two copies (Figure 2b), all localized in a single genomic context (Figure 3c). A similar observation is made for the PLT5 and AIL1 subclades across all angiosperms. Therefore, it seems that genes in these subclades experience a tight regulation of gene dosage and transcription. Overexpression of Arabidopsis PLT5 was shown to induce formation of somatic embryos and enlargement of floral organs (Nole-Wilson et al., 2005; Tsuwamoto et al., 2010). Similarly, Arabidopsis plt1 plt2 double mutants are disrupted in root meristem identity and develop short roots, whereas PLT2 overexpression induces ectopic formation of root meristems (Aida et al., 2004; Galinha et al., 2007). No function has yet been attributed to AIL1. The extreme conservation of the AIL1 subclade creates an extra incentive to
characterize its presumptive key role functionally in plant development.

For other subclades, variation in copy number does not reflect variation in synteny. ANT genes vary largely in copy number, but only belong to a single syntelog cluster (Figures 2b and 3c). This can be explained in two ways, i.e. (i) flexibility in ANT gene dosage, but restriction in regulation, or (ii) subfunctionalization of ANT genes restricting variation in expression. Conversely, the WR1/3/4-L HMMERlog copy number is more or less constant across non-Brassicaceae lineages with one or two copies (Figure 2b), but WR1/3/4-L HMMERlogs belong to four syntelog clusters (Figure 3c). Copy number variation in the PLT3/7 subclade does not differ substantially between monocots and eudicots (Figure 2b). However, the genomic context of PLT3/7 subclade is unique between both lineages (Figure 3c), suggesting an evolutionary divergence in regulatory networks of phyllotaxis and rhizotaxis. In contrast, the PLT5 subclade that also contributes to phyllotaxis and rhizotaxis (Prasad et al., 2011; Hofhuis et al., 2013) is strongly conserved (Figure 3c). This suggests that lateral organ positioning has an ancestral component through PLT5, and an eudicot/monocot-specific regulatory component through PLT3/7. In conclusion, we show that combining phylogenetic analysis and syntelog clustering is a powerful tool to assess conservation and lineage specificity of individual subclades within angiosperm gene families. This serves as a valuable resource for linking trait evolution to specific genomic events.

EXPERIMENTAL PROCEDURES

Identification of HMMERlogs in 107 angiosperm species

Proteomes of 107 angiosperm species were retrieved from public repositories as described by Zhao and Schranz (2019). Sequence similarity searches were performed using the AP2 alignment in Stockholm format (AP2, 2×PF00847; EREBP, 1×PF00847) using HMMER v.3.2.1. (El-Gebali et al., 2018). Hits below the default inclusion threshold (E < 0.01) were extracted. Protein sequences containing two AP2 domains were classified as AP2 subfamily members, and those with a single AP2 domain as EREBP subfamily members. Proteins with more than two detected AP2 domains were also assigned as AP2 members to account for erroneously recognized repeated domains and misannotations, although the potential existence of these proteins warrants experimental verification (Table S4). Protein sequences lacking a start codon were filtered out (Data S1). BLASTP analysis of the ambiaAP2 subclade was performed with default settings against the A. thaliana non-redundant (nr) protein database (taxid: 3702; Altschul et al., 1990). HMMERlogs containing NFYB domains (PF00808) or single B3 domains (PF02362) were identified in a similar way as described above.

Multiple sequence alignment and phylogenetic analysis

Full-length protein sequences were aligned with MAFFT v7 using the FFT-NS-2 progressive algorithm with a gap penalty of 1.0 (Katoh et al., 2017). Sparsely sized sequences were filtered by length range (AP2, 200–800 amino acids; EREBP, 100–525; NFYB 100–350; and B3, 100–1250 respectively). Gapped positions in filtered multiple sequence alignments were removed by trimAl (Capella-Gutiérrez et al., 2009). For AP2 and EREBP sequences, the automatic ‘-gappyout’ mode was used, which retained several hundred positions in the multiple sequence alignments, including the AP2 domain(s). This setting proved to be unsuitable for trimming the more variable NFYB and B3 proteins. Instead, approximately 300 positions with the least gaps were kept by altering the ‘-gt’ and ‘-cons’ flags. Maximum-likelihood trees were constructed from trimmed alignments with IQ-TREE (Excoffier et al., 2013), using the LG substitution matrix and 1000 ultrafast bootstraps (Le and Gascuel, 2008; Nguyen et al., 2015; Hoang et al., 2018). Phylogenetic trees were edited in iTOL 4.4 (Letunic and Bork, 2019). The EREBP tree was simplified by collapsing branches supported by <500 iterations.

Synteny network extraction and syntelog clustering

Synteny networks were obtained by extracting the identified HMMERlogs and their syntenic connections from the proteome-wide angiosperm synteny network presented in Zhao and Schranz (2019). This network was generated with SynNet-Pipeline, which is available at https://github.com/zhaoao1987/SynNet-Pipeline. In this network, genes were marked as syntenic when they belonged to a 51-gene window in which at least five other genes shared homology. Residual non-HMMERlog genes were discarded. Syntenic HMMERlog genes (syntelogs) were clustered with the Info-map algorithm in R (Rosvall and Bergstrom, 2008). Redundant connections were removed, and only syntelogs with a k-core >3 were kept. Synteny networks were visualized and coloured with Cytoscape 3.7.1 (Shannon et al., 2003). Phylogenomic profiles were created by counting the number of syntenic HMMERlog genes per syntelog cluster in all 107 species. Subsequent hierarchical clustering was performed using the HEATMAP.2 package in R. Synteny information was used to define AP2 subclades and to determine syntenic connections between subclades. To eliminate one-to-many syntenic connections between subclades, a threshold was set at a maximum of 10 connections originating from a single syntelog. Lineage specificity of synteny was determined by counting the number of syntelog clusters per subclade. Syntelog clusters were considered to be lineage-specific when containing at most two species belonging to other taxa. A cluster was not assigned to a subclade when containing <10% HMMERlogs belonging to that particular subclade.

Cluster overlap index analysis

As a measure for the relative degree to which a certain gene family is conserved, we devised the cluster overlap index. This index is calculated by performing pairwise interspecies comparisons of overlapping syntelog clusters. For example, species A has a cluster overlap index of 0.8 with species B, when eight of 10 clusters contain syntelogs of both A and B. Distributions of overlap indices are a measure to compare relative conservation of synteny between AP2 and EREBP subfamilies, which is as long as the same input species are used. Angiosperm species lacking syntelogs in both families were excluded from this analysis.

ACKNOWLEDGEMENTS

We acknowledge the use of icons designed by Freepik in the graphical abstract. We thank Tao Zhao and Setareh Mohammadin for technical assistance and providing scripts. Viola Willemsen is acknowledged for constructive suggestions and ideas for improving the manuscript.

© 2020 The Authors. The Plant Journal, published by Society for Experimental Biology and John Wiley & Sons Ltd, The Plant Journal (2020), doi: 10.1111/tpj.14843
AUTHOR CONTRIBUTIONS

MHLK, MES and KB designed the research. MHLK performed experiments. MHLK and KB analyzed data. MHLK, MES and KB wrote the manuscript. All the authors read and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The AP2/EREBP data generated in this study are available at Harvard Dataverse (https://doi.org/10.7910/DVN/CXKAQA). This includes the HMMERlog protein sequences, (un)trimmed multiple sequence alignments, phylogenetic trees and in-house Bash/R scripts.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Maximum-likelihood phylogenetic tree of AP2 HMMERlogs displaying unfiltered syntenic relationships within and between subclades.

Figure S2. Lineage-specific synteny of a LEC1/L1L cluster in Brassicaceae.

Figure S3. Maximum-likelihood phylogenetic tree of EREBP HMMERlogs displaying syntenic relationships within and between subclades.

Figure S4. Copy number variation in EREBP subclades in angiosperms.

Figure S5. Phylogenomic profile of EREBP syntenogs in 107 angiosperm species.

Figure S6. Bar plot displaying the number of syntelog clusters and total number of HMMERlogs per EREBP suboclade.

Table S1. List of identified HMMERlogs per species and per AP2 suboclade.

Table S2. List of identified HMMERlogs per species and per EREBP suboclade.

Table S3. Best BLASTP hits of ambiaAP2 HMMERlogs against the Arabidopsis proteome.

Table S4. List of AP2 HMMERlogs with more than two AP2 domains according to HMMER and SMART predictions.

Data S1. AP2 and EREBP HMMERlogs identified by our pipeline.

Data S2. Synteny connections between AP2 and EREBP HMMERlogs.

REFERENCES

Aida, M., Beis, D., Heidstra, R., Willemsen, V., Bilou, I., Galinha, C., Nus-saume, L., Noh, Y.S., Amasino, R. and Scheres, B. (2004) The PLETHORA genes mediate patterning of the Arabidopsis root stem cell niche. Cell, 119(1), 109–120.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215(3), 403–410.

Artur, M.A.S., Zhao, T., Ligerink, W., Schranz, M.E. and Hilhorst, H.W.M. (2019) Dissecting the genomic diversification of late embryogenesis abundant (LEA) protein gene families in plants. Genome Biol. Evol. 11(2), 459–471.

Barker, M.S., Li, Z., Kidder, T.I., Reardon, C.R., Lai, Z., Oliveira, L.O., Scasci-telli, M. and Rieseberg, L.H. (2018) Most Compositeae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calcarceae. Am. J. Bot. 103(7), 1203–1211.

Baud, S., Mendoza, M.S., To, A., Harsoet, E., Lepinec, L. and Dubreucq, B. (2007) WRINKLED1 specifies the regulatory action of LEAFY COTYLEDON2 towards fatty acid metabolism during seed maturation in Arabidopsis. Plant J. 50(5), 825–838.

Capella-Gutierrez, S., Silla-Martinez, J.M. and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972–1973.

Cernac, A. and Benning, C. (2004) WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in Arabidopsis. Plant J. 40(4), 576–585.

Clark, J.W. and Donoghue, P.C.J. (2018) Whole-genome duplication and plant macroevolution. Trends Plant Sci. 23(10), 933–945.

Conflonlieri, M., Carelll, M., Galimberti, V., Maceovi, A., Panara, F., Biggi-giarelli, M., Scotti, C. and Calderoni, O. (2014) Seed-specific expression of AINTEGUMENTA in Medicago truncatula led to the production of larger seeds and improved seed germination. Plant Mol. Biol. Rep. 32(5), 957–970.

Dewey, C.N. (2011) Positional orthology: putting genomic evolutionary relationships into context. Brief. Bioinform. 12(5), 401–412.

Dodsworth, S., Chase, M.W. and Leitch, A.R. (2016) Is post-polypliodization diploidization the key to the evolutionary success of angiosperms? Bot. J. Linn. Soc. 180(1), 1–5.

El-Gebali, S., Mistry, J., Bateman, A. et al. (2019) The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1), D427–D431.

Elliott, R.C., Betzner, A.S., Huttner, E., Oakes, M.P., Tucker, W.Q., Gerentes, D., Perez, P. and Smyth, D.R. (1996) Arabidopsis – a genomic resource for plant plant macroevolution. Plant Mol. Biol. 32(5), 957–970.

Galinha, C., Hofhuis, H., Luijten, M., Willemsen, V., Bilou, I., Heidstra, R. and Scheres, B. (2007) PLETHORA proteins as dose-dependent master regulators of Arabidopsis root development. Nature, 449(7165), 1053–1057.

Guo, Y.L. (2013) Gene family evolution in green plants with emphasis on the origin and evolution of Arabidopsis thaliana genes. Plant J. 73 (6), 941–951.

Hoang, D.T., Chernomor, O., Haeseleer, A.V., Minh, B.Q. and Vinh, L.S. (2018) UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35(2), 518–522.

Hofhuis, H., Laskowski, M., Du, Y., Prasad, K., Grigg, S., Pinon, V. and Scheres, B. (2013) Phyllotaxis and rhizotaxis in Arabidopsis are modified by three PLETHORA transcription factors.Curr. Biol. 23(11), 956–962.

Horstman, A., Li, M., Heidmann, I., Weemen, M., Chen, B., Muino, J.M., Angenent, G.C. and Boutilier, K. (2017) The BABY BOOM transcription factor activates the LEC1–BABIUS3–LEC2 network to induce somatic embryogenesis. Plant Physiol. 176(2), 848–857.

Horstman, A., Willemsen, V., Boutilier, K. and Heidstra, R. (2014) AINTEGU-MENTA-LIKE proteins: hubs in a plethora of networks. Trends Plant Sci. 19(9), 146–157.

Jiao, Y. and Paterson, A.H. (2014) Polypliody-associated genome modifications during plant land plant evolution. Philos. Trans. R Soc. B Biol. Sci. 369 (1648), 20130355.

Jojuku, K.D., den Boer, B.G.W., Montagu, M.V. and Okamura, J.K. (1994) Control of Arabidopsis flower and seed development by the homeotic gene APETALA2. Plant Cell, 6(9), 1211–1225.

Katoh, K., Rozewicki, J. and Yamada, K.D. (2017) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. bbx108. 20(4), 1100–1106.

Kim, S., Soltis, P.S., Wall, K. and Soltis, D.E. (2006) Phylogeny and domain evolution in the APETALA2-like gene family. Mol. Biol. Evol. 23(1), 107–120.

Krizek, B.A. (1999) Ectopic expression of AINTEGUMENTA in Arabidopsis plants results in increased growth of floral organs. Dev. Genet. 25(3), 224–236.

Krogan, N.T., Hogan, K. and Long, J.A. (2012) APETALA2 negatively regulates multiple floral organ identity genes in Arabidopsis by recruiting the
Evolutionary divergence of APETALA2 transcription factors

Prasad, K., Grigg, S.P., Barkoulas, M. et al. (2011) Arabidopsis PLETHORA transcription factors control phytolaxis. Curr. Biol. 21(13), 1123-1128.

Rashid, M., Guanyuan, H., Guangxiao, Y., Hussain, J. and Xu, Y. (2012) AP2/ERF transcription factor in rice: Genome-wide canvas and syntenic relationships between monocots and eudicots. Evol. Bioinform. Online, 8 (11), 321-355.

Riechmann, J.L. and Meyerowitz, E.M. (1998) The AP2/EREBP family of plant transcription factors. Biol. Chem. 379(6), 633-646.

Rosvall, M. and Bergstrom, C.T. (2008) Maps of random walks on complex networks reveal community structure. Proc. Natl Acad. Sci. USA, 105(4), 116-1123.

Ruelens, P., de Maagd, R.A., Proost, S., Theissen, G., Geuten, K. and Kaufmann, K. (2013) FLOWERING LOCUS C in monocots and the tandem origin of angiosperm-specific MADS-box genes. Nat. Commun. 4, 2289.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13(11), 2498-2504.

Shigyo, M., Hasebe, M. and Ito, M. (2006) Molecular evolution of the AP2 subfamily. Gene, 366(2), 256-265.

Smith, Z.R. and Long, J.A. (2010) Control of Arabidopsis apical-basal embryo polarity by antagonistic transcription factors. Nature, 464(7287), 423-426.

Song, X., Li, Y. and Hou, X. (2013) Genome-wide analysis of the AP2/ERF transcription factor superfamily in Chinese cabbage (Brassica rapa ssp. pekinensis). BMC Genom., 14(1), 573.

Steven, P.F. (2001 onwards) Angiosperm Phylogeny Website. Version 14, July 2017.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2020) Synteny and collinearity in plant genomes. Science, 320(5875), 498-498.

To, A., Joubes, J., Barthole, G., Lecureuil, A., Scagnelli, A., Jasinski, S., Lepineac, L. and Baud, S. (2012) WRINKLED transcription factors orchestrate tissue-specific regulation of fatty acid biosynthesis in Arabidopsis. Plant Cell, 24(12), 5007-5023.

Trusba, S., Isnard, S., Barthelemy, D. and Olson, M.E. (2016) Trait coordination, mechanical behaviour and growth form plasticity of Amborella trichopoda under variation in canopy openness. AoB Plants, 8(1), p58088.

Tsuyamoto, R., Yokoi, S. and Takahata, Y. (2010) Arabidopsis EMBRYO-MAKER encoding an AP2 domain transcription factor plays a key role in development and change from vegetative to embryonic phase. Plant Mol. Biol. 74(4-5), 481-492.

Wang, L., Ma, H. and Lin, J. (2019) Angiosperm-wide and family-level analyses of AP2/ERF genes reveal differential retention and sequence divergence after whole-genome duplication. Front. Plant Sci. 10(1), 196.

Wang, F. and Perry, S.E. (2013) Identification of target transcripts of FUSCA3, a key regulator of Arabidopsis seed development. Plant Physiol. 161(3), 1251-1264.

Xiao, Q., Li, G., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S. and Paterson, A.H. (2019) Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. BMC Genome Biol. 20, 38.

Zhang, B., Wang, L., Zeng, L., Zhang, C. and Ma, H. (2015) Arabidopsis TOE proteins convey a photoperiodic signal to antagonize CONSTANS and regulate flowering time. Genes Dev. 29(9), 975-987.

Zhao, T., Holmer, R., de Bruijn, S., Angenent, G.C., van den Burg, H.A. and Schnitz, M.E. (2017) Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. Plant Cell, 29(6), 1278-1292.

Zhao, T. and Schranz, M.E. (2017) Network approaches for plant phylogenomic synteny analysis. Curr. Opin. Plant Biol. 36(1), 129-134.

Zhao, T. and Schranz, M.E. (2019) Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. Proc. Natl Acad. Sci. USA, 116(6), 2165-2174.

Zhu, Q.H. and Hellwig, F.A. (2011) Regulation of flowering time and floral patterning by miR172. J. Exp. Bot. 62(2), 487-495.

Zumajo-Cardona, C. and Fabian-Mora, N. (2016) Evolution of the APETALA2 gene lineage in seed plants. Mol. Biol. Evol. 33(7), 1818-1832.