Phylogenomic synteny network analyses reveal ancestral transpositions of auxin response factor genes in plants

Bei Gao1, Liuqiang Wang2, Melvin Oliver3, Moxian Chen*4 and Jianhua Zhang1,5*

Abstract

Background: Auxin response factors (ARFs) have long been a research focus and represent a class of key regulators of plant growth and development. Integrated phylogenomic synteny network analyses were able to provide novel insights into the evolution of the ARF gene family.

Results: Here, more than 3500 ARFs collected from plant genomes and transcriptomes covering major streptophyte lineages were used to reconstruct the broad-scale family phylogeny, where the early origin and diversification of ARF in charophytes was delineated. Based on the family phylogeny, we proposed a unified six-group classification system for angiosperm ARFs. Phylogenomic synteny network analyses revealed the deeply conserved genomic syntenies within each of the six ARF groups and the interlocking syntenic relationships connecting distinct groups. Recurrent duplication events, such as those that occurred in seed plants, angiosperms, core eudicots and grasses contributed to the expansion of ARF genes which facilitated functional diversification. Ancestral transposition activities in important plant families, including crucifers, legumes and grasses, were unveiled by synteny network analyses. Ancestral gene duplications along with transpositions have profound evolutionary significance which may have accelerated the functional diversification process of paralogues.

Conclusions: The broad-scale family phylogeny in combination with the state-of-art phylogenomic synteny network analyses not only allowed us to infer the evolutionary trajectory of ARF genes across distinct plant lineages, but also facilitated to generate a more robust classification regime for this transcription factor family. Our study provides insights into the evolution of ARFs which will enhance our current understanding of this important transcription factor family.

Keywords: Auxin, ARF, Transcription factor, Gene duplication, Genomic synteny

Background

The plant hormone auxin, or indole-3-acetic acid, controls many physiological and developmental processes in land plants including but not limited to organogenesis, tissue differentiation, apical dominance, gravitropism as reviewed previously [1, 2]. Completion of the genomes of the moss Physcomitrella patens, [3, 4], the liverwort Marchantia polymorpha [5] and the lycophyte Selaginella moellendorffii [6] revealed that many core functional proteins required for auxin biosynthesis, perception, and signaling were present in the early-diverging land plant lineages. Comprehensive evolutionary studies also suggested that the auxin molecular regulatory network evolved at the latest in the common ancestor of embryophytes [7, 8] and the auxin response factor (ARF) genes evolved from...
the charophyte ancestors [9]. A recent review demonstrated that with the exception of the PGP/ABCB auxin transporters, homologues of all the other core components for hormonal control of physiology and development by auxin could be identified in P. patens [10]. Changes in auxin perception and signaling that occurred through evolution could have generated the diversification of plant forms that occurred during the past \( \sim 474–515 \) million-year history of the land plants [11], which eventually led to the complex vegetative innovations that shape the modern terrestrial and freshwater ecosystems [2].

Auxin response factors, as core components in auxin signaling, have long been a focus of plant signaling research [12]. The 23 ARFs identified in the Arabidopsis thaliana genome were phylogenetically clustered into three subfamilies (Clades A, B and C) which were subsequently divided into seven groups (ARF9, ARF1, ARF2, ARF3/4, ARF6/8, ARF5/7 and ARF10/16/17), a classification that was well supported by ARF genes from other angiosperms and representative non-flowering lineages [13]. Generally, ARF proteins can be functionally divided into transcriptional activators (ARF5–8 and 19 in A. thaliana) and repressors (remaining ARFs in A. thaliana) with well-characterized functional domain architectures [13, 14]. ARFs bind to the auxin response elements (AuxRE: TGTCTC) in the promoter region of downstream auxin-inducible genes [15] and function in combination with Aux/IAA repressors, which dimerize with ARF activators in an auxin-regulated manner [14, 16]. Unlike ARF activators, few reports have demonstrated that ARF repressors are able to interact with other ARF proteins or Aux/IAA proteins [17]. Recent work revealed a newly identified mechanism whereby the IAA32 and IAA34 transcriptional repressors are stabilized by the transmembrane kinase 1 (TMK1) at the concave side of the apical hook of the kinase to regulate ARF gene expression and ultimately inhibit growth [18].

In most of the well-established transcription factor annotation procedures, such as those implemented by the PhnTFDB [19], PlantTFDB [20], iTAK [21] and TAPScan [22], ARFs were identified using two signature domains: the B3 (PF02362) domain and the auxin-response (PF06507) domain, although some ARF proteins (e.g. ARF23 in A. thaliana) may be truncated and lack the auxin-response domain [13, 14, 23]. Finet et al. [13] established a robust and comprehensive phylogenetic framework for the ARF gene families, however ARF genes from non-flowering plants were under-represented. Comprehensive annotation of transcription factors covering distinctive plant clades demonstrated that a number of plant specific transcription factor families (including ARF) evolved in streptophytic algae [9, 22, 24], suggesting an earlier origin of ARF than that proposed by Finet and colleagues [13].

Compared to conventional gene family studies that focus on one or a limited number of species of interest [25–27], phylogenetic studies on a broader scale that include multiple plant lineages were able to generate more robust insights into the evolutionary process that gave rise to the modern assemblage of a target gene family [13, 28]. The inclusion of genomic synteny data provides important information that impacts the determination of the evolutionary past of a gene family, especially when the gene family of interest evolved in parallel with ancestral genome duplication events [29]. The conventional genomic block alignment that connects orthologues, retained on genomic syntenic blocks, worked well for a limited number of species [29, 30], but a network approach was more effective when multiple genomes were included in the synteny analyses [31, 32]. A comprehensive genomic synteny network can be constructed using nodes to represent the target genes and associated adjacent genomic blocks and the network edges (connecting lines) to represent syntenic relationships [29, 32]. The recently established phylogenomic synteny network methodology was able to integrate and summarize genomic synteny relationships to uncover and place genomic events (e.g. ancient tandem duplications, lineage-specific transposition activities) into the evolutionary past of a target gene family [31, 32].

In this study, we collected more than 3500 ARF members to generate a comprehensive gene-family phylogeny with the aim of filling evolutionary gaps in the non-flowering plants and splitting the long branches present in the current phylogeny [13]. We propose an updated model for the evolution of ARF family that covered the major streptophytic clades that was based on the six-group classification system we proposed for the ARF genes in angiosperms. Phylogenomic synteny network analyses of angiosperm genomes revealed the deep positional conservation of ARF members within each of the six groups. Detailed individual synteny network analyses together with phylogenetic reconstructions for the six ARF groups revealed their distinctive evolutionary histories. Ancestral duplication events in angiosperms, and subsequent WGDs in eudicots and monocots have contributed to the expansion of ARF members. Ancestral lineage-specific transpositions in important angiosperm families such as cucifers, legumes and grasses were also unveiled. Together, the results presented here add to our current understanding of the evolutionary process that established ARF genes in plants. We also expect this broad-scale evolutionary framework could help direct future functional studies that further explore the interplay between auxin signaling and the evolution of land plants.
Results and discussion

Evolution of auxin response factors in streptophytic algae

To generate a broad-scale phylogenetic profile for ARF genes in plants, we collected a total of 3502 ARF homologues in the streptophytes. ARFs were present in all major clades of streptophytes including charophytes, hornworts, liverworts, mosses, lycophytes, ferns and seed plants (Additional file 1: Fig. S1). In charophytes, the Auxin-response domain was not detected although some charophyte genes did contain the B3 domain. Genes containing both the B3 (PF02362) and the Auxin-response (PF06507) domains were identified in streptophytic algae (charophytes). This was consistent with the observation that a number of plant-specific transcription factors evolved in streptophytic algae [22]. Charophytes represented a paraphyletic clade encompassing successive sister lineages to the land plants [22, 33]. We identified ARF homologues in species that are found in three charophyte orders: Zygnematales, Coleochaetales and Chlorokybales, but ARF homologues were not identified in the transcriptomes of Charales and Klebsormidiales. However, the presence of ARF in charophytes was affirmed by the Chara braunii (Charales) genome [24]. The identification of the single ARF gene in Chlorokybus atmosphyicus (Chlorokybaceae) suggests that the origin of ARF genes probably trace back to the root position of streptophytes (Additional file 1: Fig. S1). This observation suggests an earlier origin of ARF gene than those reported earlier [13, 22] and consistent with a previous study [8].

Broad-scale phylogenetic profile of ARFs in plants

Overall, the numbers of ARF genes in individual angiosperm genomes are greater than those in the individual genomes of non-flowering plants and the ‘recent’ polyploids, such as Glycine max, possess conspicuously more ARF genes than other plants (Additional file 1: Fig. S2). The inclusion of homologues identified from the 1KP transcriptome database provided a comprehensive atlas for the ARF family phylogeny. Overall, the broad-scale phylogeny of ARFs generated in this analysis was closely in parallel with the phylogenetic relationships among plant lineages (Fig. 1) derived from large-scale phylotranscriptomic study [34]. The phylogenetic tree generated from the ARF gene collection was rooted by the ARF gene from Chlorokybus atmosphyicus, an early diverging charophyte, and exhibited a consistent tree topology with that reported previously [13]. The incorporation of transcriptomic data from non-flowering plants enabled long evolutionary branches to be split. The phylogenetic analyses also provided robust evidence that angiosperm ARFs could be separated clearly into three major subfamilies (Clade A, B and C; consistent with previously reported groupings [13]). The three major subfamilies encompassed the six groups (designated as Group 1 through VI) in this study (Fig. 1). In comparison to the classification proposed by Finet et al. [13], the Group-I ARFs contained the ARF1 and ARF9 subfamilies (which were likely to have derived from an ancient angiosperm-wide duplication) and Group II through VI correspond to the ARF 2, ARF 3/4, ARF 6/8, ARF 5/7 and ARF 10/16/17 subfamilies (Additional file 1: Table S1), respectively.

Groups I, II and III were clustered in the subfamily clade-B, groups IV and V in clade-A and group VI in clade-C. The reconstructed family phylogeny suggested clade-C as a sister group to clades A and B. ARFs from the charophytes (Zygnematales and Coleochaetales) were separated into clade-C and clade-B failing to partition into a basal mono- or para-phyletic clade, which suggested an ancient diversification of ARF genes within the charophytes. The family phylogeny also revealed that each of the six angiosperm ARF family groups were located with gymnosperm ARF genes as the closest sister lineage. The tree branches of gymnosperm ARF genes are conspicuously shorter than those for angiosperms (Fig. 1 and Additional file 1: Fig. S3), which suggested lower amino acid substitutional rates and higher levels of protein sequence conservation in gymnosperm ARF genes, likely a result of longer generation times that are common in the gymnosperms [35].

Clade-A contains ARF genes that cover all major embryophyte clades and contains ARF genes of group-III together with orthologues from gymnosperms and ferns. The ARF genes from seed plants and ferns were separated into two major clades which are sister to each other which constituted a tree topology that was consistent with two child clades derived from an ancient duplication. While lycophyte ARF genes were placed outside of and sister to the large duplication clade shared by ferns and seed plants. ARF genes from hornworts were identified as basal-most in clade-A, followed by genes from mosses and liverworts.

Clade-B was the most diversified lineage containing the angiosperm group I and II genes and along the gymnosperm orthologous genes delineated a conspicuous seed-plant duplication (the event) [36]. However, ARF genes from hornworts, liverworts and ferns were mixed into this large duplication clade (Fig. 1). We hypothesize that they might be derived from convergent evolution, though the possibilities of horizontal gene transfer or sequence contaminations cannot be eliminated. Genes from ferns, mosses, liverworts and lycophytes were placed as successive sister lineages to this duplication clade.

Clade-C was situated as the basal clade with a relatively simple phylogenetic profile and contains genes from every major plant lineage (from charophytes to...
angiosperms, Fig. 1). This configuration updated the evolutionary model in which clade-C ARFs were absent in gymnosperms [13].

The broad-scale phylogenetic analyses in this study established a robust and unified six-group classification system for angiosperm ARF genes, which is consistent with previous phylogenetic and domain architecture studies [13, 14]. The relative phylogenetic positions of other land plant lineages were also clarified (Fig. 1), providing a consistent phylogenetic framework for subsequent synteny network analyses.

**Evolutionary trajectory of ARFs augmented with current genomic and transcriptomic data**

In concordance with the phylogenetic analyses described by Finet et al. [13], we augmented the evolutionary trajectory of ARF family in plants with gene sequences from the currently available genomic and transcriptomic data. The resulting phylogenetic trajectory path (Fig. 2) suggested that the three ARF subfamilies (clades A, B and C) were likely diversified through an ancient duplication in the charophytes, which is consistent to the evolutionary trajectory proposed previously [7, 8]. Tree uncertainties and unresolved land plant phylogenies were also reflected in the ARF gene-family phylogeny, leaving some of the evolutionary processes elusive.

All of the ARF transcriptional activators (ARF 5–8 and 19 in A. thaliana) were clustered in the clade-A subfamily. Within clade-A the ARF genes were well-conserved in all land plant lineages and appear to have experienced a conspicuous ancient duplication event that occurred in the ancestor of ferns and seed plants. This ancient duplication generated groups IV (ARF6 and 8 in A. thaliana) and V (ARF5, 7 and 19 in A. thaliana) in the angiosperms and the corresponding sister groups in gymnosperms and ferns (Fig. 2). The ARF genes in bryophytes (including hornworts, liverworts and mosses) and lycophytes were outside of this duplication. The ARF genes in clade-A also exhibited a gene tree topology consistent with the ‘hornwort-sister’ land
plant phylogeny in contrast to the ‘bryophytes-monophyletic’ phylogeny [11, 34]. The evolutionary well-conserved aspect of the ARF activator genes indicates an early genetic foundation for auxin signaling networks in the embryophytes [10].

In clade-B, unlike clade-A, ARF genes from hornworts were not found densely populated at the basal position of the subtree and some hornwort ARFs were found clustered with angiosperm ARF genes, making the evolution of clade-B ARFs in hornworts elusive. The trajectory analysis suggests two ancient duplications in clade-B, an embryophyte duplication shared by mosses, liverworts and tracheophytes that occurred before the diversification of groups I/II and group III, and a seed plant duplication that generated the groups: I and II. However, the close sister groups for group III were only found in the gymnosperms and ferns which suggested there were gene losses in mosses, liverworts and lycophytes (Fig. 2).

The subfamily of clade-C is well-conserved, covering all streptophytic lineages, and generated the simplest phylogenetic profile (Fig. 2), containing the group-VI angiosperm ARF genes (the ARF 10, 16, and 17 in A. thaliana). Hornwort ARF genes were placed as direct sisters to the vascular plants (tracheophytes) and the ARF genes of mosses and liverworts were placed at the base of the subtree (Fig. 1), generating discrepancies in the gene tree topology and the phylogeny of early land plant lineages.

**Phylogenomic synteny network analyses of ARF genes**
The broad-scale phylogenetic analyses suggested some subtree topologies that are consistent with the occurrence of ancient gene duplications but genomic synteny analyses are required to provide more substantive evidence [37]. The recently established synteny network approach, taking advantages of accumulated plant genomes, was able to provide such substantive evidence for ancient evolutionary processes of a specific gene family [31, 32]. Applying this approach, we conducted a phylogenomic syntenic network analyses for ARF genes using a collection of available plant genomes (Additional file 1: Fig. S2). Syntenic ARF genes (syntelogs) were observed in some non-flowering plants (e.g. a lycophyte and a moss), but all represented in-paralogues which were considered to have derived from lineage-specific duplications. The ARF genes identified in angiosperm genomes were the primary target of the analysis and used as anchors to construct the genomic synteny network.

Among the 1227 annotated angiosperm ARF genes containing valid B3 and Auxin-response domains (Additional file 2: Table S2), 1096 (89.3%) were detected to be located within genomic synteny regions that demonstrated genomic collinear relationships with at least one other ARF gene, and a total of 18,511 syntenic connections among ARF genes were detected (Fig. 3a, b). Consistent with the family phylogeny described...
previously, most of the genome synteny connections were observed within each of the six groups. ARF genes from distinctive ARF groups were syntenically connected (Fig. 3a), for example, ARF genes from group VI were connected to ARF genes from group III/IV/V and group III ARF genes with group I. The ARF synteny network analyses uncovered a total of 82 inter-group connections (Fig. 3a).

In the ARF gene synteny network, we detected 96 ARF genes that did not pass our ARF identification procedure but were demonstrated to be homologous and syntenic to the annotated ARF genes. These syntelogs were further inspected and most contained truncated B3 and/or Auxin-response domains or lacking either or both of these signature domains. These truncated or pseudogenes that were retained in the syntenic genomic blocks were not incorporated in the phylogenetic analyses, however, we were able to assign and label them into one of the six angiosperm ARF gene groups by aligning them to classified angiosperm genes. In this way, both intact (total 1096) and truncated (total 96) ARF genes involved in the synteny network were classified. The classification for these truncated genes was considered reliable because of the distant phylogenetic relationships among the six groups (Fig. 1). This may suggest that using genomic syntenic relationships could be a robust approach for detecting pseudogenes retained in the syntenic genomic blocks and which exhibit significant local sequence identity with intact functional paralogues.

ARF genes from each group were found in separate and distinct syntenic communities in the initial synteny network visualization (Fig. 3b). The ARF synteny network was further dissected to find subnetwork communities by the use of clique percolation clustering at k = 3 implemented in CFinder v2.0.6 [38]. A total of 25 communities (numbered 0 through 24) (nodes clustered within a subnetwork usually possess more connections in its community than with nodes in other communities) were obtained (Fig. 4). Among the 1192 ARF syntelogs that were extracted from the synteny network database, 1128 (94.6%) were identified in the 25 network communities, other syntelogs that had a single syntenic connection or were not involved in a clique (at k = 3) were excluded. For example, among the 22 ARF genes in Arabidopsis thaliana, 17 members were clustered in 13 synteny network communities (Fig. 4a). The chromosome-level genome assemblies represented the best material for genome synteny analyses, but some plant genome assemblies currently available are still highly fragmented. For example, in the Malus domestica (apple) genome, only one ARF gene was clustered in the synteny network because the genome assembly version we obtained from Phytozome database and that was used in our synteny network construction was fragmented (approximately 881.3 Mb.
arranged in 122,107 scaffolds) (Fig. 4a). However, the network approach using multiple plant genomes appeared to be error-tolerant and the results were unaffected by the inclusion of a few fragmented genomes [31].

Species compositions for each of the 25 synteny network communities (Fig. 4a) indicate that network communities 4 and 5 are angiosperm-wide, containing ARF genes from monocots, eudicots and Amborella, Community 23, on the other hand, only contains ARF genes from monocots and community 24 is solely confined to ARF genes from eudicots. Other communities are lineage-specific such as community 21 which only contains ARF genes from grasses, communities 13 and 14 that are specific to legumes, and communities 16 and 22 that are specific to the genus Brassica.

Subnetwork communities were separately visualized, using node colors to depict different plant lineages and node shapes, to delineate ARF genes from the different classification groups (Fig. 4b). Community 0 (labeled as ‘VI-16-45’) consisted of ARF members from group-VI, with a total of 16 nodes and 45 connections within the community. Some syntenic communities contained ARF genes from multiple groups. Community 5 was recognized as the largest community with 226 nodes and 4742 connections, and nodes in this community were primarily ARF genes from group-VI and group-IV, with a minority of members from group-III (3 nodes) and group-V (1 node). The mixed group communities suggest the existence of ancient tandem duplications [31], where duplicated paralogues were likely lost in the ancestral genome such that ancient tandem paralogues are not seen in most current plant genomes, but synteny network analyses reflect them as multigroup communities. Consistent with this hypothesis, tandem ARF genes from distant groups were rarely present in the genome of a single species used in the analysis (Additional file 1: Table S1). To illustrate this, the ARF gene (scaffold00029187) from Amborella was classified as a member of group-IV, but it had a syntenic connection with group-VI ARF genes from Oryza sativa (LOC_Os10g33940), Oropetium thomaeum (Oropetium_20150105_02810A) and Phaseolus vulgaris (Phvul_003G075800). This could be explained by the occurrence of an ancient tandem gene duplication that was generated prior to the separation of groups VI and IV. Following the speciation of basal angiosperms and eudicots plus monocots, the group-VI member was lost in Amborella, and the group-IV member was lost in the ancestor of monocots and eudicots resulting in the syntenic relationship seen between group-VI and group-IV ARF genes. The inter-group genomic syntenic
connections not only provided evidence for ancient gene duplications followed by lineage-specific gene losses, but also suggested that modern ARF genes evolved from a common ancestor present in the streptophytes.

**Evolutionary characteristics for each of the six groups of ARFs in angiosperms**

The global phylogenetic and synteny network analyses generated a robust six-group classification system for ARF genes and indicated pervasive intra-group synteny relationships. To elaborate the evolutionary processes within each of the six groups, individual phylogenetic trees for angiosperm genes in each of the six groups were estimated separately and syntenic connections within each network community were mapped onto the six gene trees [39]. Along with the ARF genes identified from Phytozome plant genomes, the ARF genes from basal ANA grade angiosperms and magnoliids were incorporated in the phylogenetic analyses, however most of these ARF gene sequences were derived from transcriptomes and thus did not provide syntenic information. The number of angiosperm (including eudicots, monocots, magnoliids and ANA grade) ARF genes in each of the six groups ranged from 190 (group II) to 318 (group I). Below we describe the primary evolutionary characteristics for the six ARF groups separately.

**Group-I**: This group represented the largest clade (containing 318 angiosperm ARF gene members) of the six groups (Fig. 5a). An evident angiosperm-wide duplication (delineated as groups IA and IB) was identified from the tree topology with the three relevant bootstrap values supporting the duplication branch and the two child clades greater than 95%. Both IA and IB clades include genes from monocots, eudicots, magnoliids and basal angiosperm lineages. The single ARF gene member from *Amborella* was placed as sister to both the IA and IB duplication clades, suggesting that the ARF

![Fig. 5](image-url)
gene duplication likely occurred after the separation of *Amborella* from other angiosperms.

Network communities associated with group-I included angiosperm-wide communities 4 (116 nodes) and 2 (65 nodes) (Fig. 4b), which align to groups IA and IB (Fig. 5a), respectively. Group IA was consistent with the designation ARF9 and group IB with the designation ARF1 in *A. thaliana* as reported previously [13]. The number of *ARF* genes included in group IA was greater than in group IB, particularly for the *ARF* genes from superrosids. The core-eudicot duplication [40] may have contributed to the expansion of the family, but some *ARF* genes from the magnoliids were included in the duplication clade with a bootstrap supporting value for the duplication node lower than 70%. In addition, some lineage-specific network communities for *ARF* genes in group-I were observed: communities 10, 11, 13, 14 and 16 are small communities containing the *ARF* genes from superrosids (Fig. 4b) and rendered as monophyletic clades in the phylogenetic analyses (Fig. 5a). The species composition analysis (Fig. 4a) for these lineage-specific communities indicated ancestral transposition activities in the Brassicaceae (communities 10, 11 and 16) and legumes (communities 13 and 14).

**Group-II:** Group-II was the smallest group (190 angiosperm *ARF* genes) and synteny network analyses revealed two main communities, 19 and 8, as depicted in Fig. 5b. Community 8 contains 26 nodes with *ARF* genes from only eudicots and *Amborella*, clustered with a group of magnoliid genes, that formed a paraphyletic clade at the basal position. While the nodes in community 19 were angiosperm-wide, and *ARF* genes from grasses were separated into two clades, one clade following the *ARF* genes in community 8 and the other clustered with the other monocots. However, the *ARF* genes clustered in each of the two grass clades did not share syntonic connections (Fig. 5b), and the two basal species (*Aquilegia* and *Amborella*) were included in both communities (Fig. 4b), suggesting the genome context (e.g. regulatory elements and adjacent genes) were altered for the *ARF* genes in the two communities. The nodes clustered in community 19 may correspond to an ancient tandem duplication in the ancestor of angiosperms as a clade of *ARF* genes from the grasses were evidently separated from other nodes, indicative of more intra than inter connections (Fig. 4b).

**Group-III:** Group-III contains 216 *ARF* genes incorporated into network communities 17 and 24 in the synteny network analyses (Figs. 4b and 5c). The phylogenetic profile for group-III genes identified them as forming two well-separated monophyletic clades (delineated as IIIA and IIIB). The group-III (community 24) contains *ARF* genes from only eudicots and magnoliids and group IIIB (community 17) is angiosperm-wide and recognized as group-IIIIB. The species composition analysis of group-III A encompassed a core-eudicot duplication shared by superrosids and superasterids, although a magnoliid *ARF* gene was clustered in this group. *ARF* genes from basal eudicots are recognized as sister to this duplication clade. Similarly, a duplication clade shared by *ARF* genes from grasses and one gene from pineapple was conspicuous and likely contributed to the generation of more *ARF* gene members in group-IIIIB in the grasses.

**Group-IV:** Group-IV contains 282 angiosperm *ARF* genes that were contained in six major network communities 5, 9, 18, 15, 3 and 6, with community 5 being the largest community and containing genes from multiple groups (Fig. 4b). Network communities 5 and 9 are angiosperm-wide and 18 contains *ARF* genes from only eudicots. The remaining three communities (3, 6 and 15) were smaller and none formed a high-confidence monophyletic clade which in turn does not support the possibility of ancestral lineage specific transpositions. By aligning the genomic synteny connections onto the phylogenetic profile, two evident clusters of *ARF* genes in this group were recognized (Fig. 5d). Communities 5, 6 and 15 were clustered into one group and communities 9 and 18 were clustered into another. Both groups were recognized as angiosperm-wide groups suggesting an angiosperm-wide duplication within group-IV, although the duplication topology cannot be easily deduced from the gene tree. In community 9, a cluster of monocot *ARF* genes contained more connections within the cluster and were separated from other nodes (Fig. 4b), suggesting the possibility of further rounds of gene duplications and losses in the evolutionary past of *ARF* genes in group IV in angiosperms.

**Group-V:** Group-V contains a total of 287 angiosperm *ARF* genes clustered in four synteny communities, 12, 20, 23 and 1 (Fig. 5e), among which communities 12 and 20 were angiosperm-wide, and communities 23 and 1 contain small numbers of monocot *ARF* genes. By integrating the synteny network and phylogenetic profile analyses, three subgroups could be identified (VA, VB and VC), and consistent with the community network analyses, the nodes in community 12 were phylogenetically separated into two subgroups (groups-VA and -VC). Nodes in communities 23 and 1 were recognized in one monophyletic clade (group-VB). An ancestral transposition in the ancestor of commelinids (including grasses, pineapple and banana genes) was evident in communities 23 and 1 (Fig. 4a, b), while an *ARF* gene from *Spirodea polyrhiza*, which is sister to commelinids, was syntenically clustered in community 20.

**Group-VI:** The group-VI included 295 *ARF* genes integrated into five syntenic network communities 5, 7, 0, 21 and 22 (Fig. 5f). Community 5 was angiosperm-wide,
community 7 encompassed primarily ARF genes from eudicot and Amborella, community 0 contained ARF genes from monocots and Amborella, and communities 21 and 22 were solely comprised of ARF genes from grasses and crucifers, respectively (Fig. 4a). Mapping the syntenic connections on the phylogenetic tree, the monophyletic clades in grasses (community 21) and crucifers (community 22) were generated and provided phylogenomic evidence for ancestral transposition activities in these two lineages. The ARF genes clustered in community 5 were phylogenetically separated into two distinct clades with some ARF genes from grasses were placed in a basal position in the group-VI phylogeny, while the nodes in community 7 were well-clustered in the family phylogeny.

In the phylogenomic syntenic network analyses we employed the maximum-likelihood gene tree generated by IQ-TREE in which more evolutionary models were implemented. We also attempted to reconstruct the ARF gene family phylogenies using RAxML (Additional file 1: Fig. S4), which generated alternative tree topologies, nevertheless, the syntenic community patterns remained constant and the major duplication clades and transposition activities could be consistently captured. Tree uncertainties may make some of the evolutionary processes that generated the ARF gene family elusive, but the syntenic network approach appears robust and uncovered evolutionary details and provided more clues for future experimental studies. For example, ARF genes were recurrently duplicated and transposed in specific lineages which suggests that the functions of these transposed genes might reveal novel regulatory elements that were captured in their altered genomic context. The transpositions that we indicated to have occurred in crucifers, legumes, commelinids and grasses were tightly associated with ancestral polyploidy events [41], which generated more possibilities in the gene regulatory network. The ancestral gene duplication together with transpositions could have greatly contributed to the expansion of the auxin regulatory network which would have had important implications in the understanding of the evolutionary processes of current land plants.

Conclusions
In this study, we generated a high-confidence broad-scale family phylogeny for ARF genes from augmented genomic and transcriptomic data, from which we summarized the evolutionary history of this focal transcription factor in streptophytes. Based on the family phylogeny, we proposed a six-group classification regime for angiosperm ARF genes. Group IV contains the ARF activators and these genes are well-conserved in all land plant clades. The Group IV subfamily phylogeny also supported the ‘hornwort-sister’ hypothesis. Genomic syntenic network analyses revealed highly conserved genomic syntenies among angiosperm ARF gene loci and within each of the six ARF gene groups. CFinder clique analyses of the ARF gene synteny network identified 25 subnetwork communities, which were further projected onto the six subfamily phylogenies. The analyses suggest that ancient duplications and transpositions have greatly contributed to the diversification of ARF genes in angiosperms. Ancestral lineage-specific transpositions of ARF genes were unveiled in crucifers, legumes, commelinids and grasses in groups I, V and VI, which were considered to have contributed to the functional diversification of gene members within a family [42]. Future studies focusing on non-angiosperm specific lineages should benefit from the evolutionary framework used in this study, especially when more genomes in these plant lineages become available [43].

Materials and methods
Collection of auxin response factors
To generate a broad-scale family phylogeny, homologues of plant ARF transcription factor genes were obtained from Phytozome v12.1.6 (https://phytozome.jgi.doe.gov/; portal.html) and the OneKP (https://db.cngb.org/onekp/) [44] databases using blastp searches filtered with an e-value threshold of 1e−5 and hmmscan with model-specific thresholds (–cut_ga). ARF gene sequences from fern genomes were collected from FernBase (https://www.fernbase.org) [45]. The protein domain composition of each of the putative ARF protein sequences were determined by querying the NCBI Conserved Domain Database [46] and only sequences that contained both definitive functional domains: B3 DNA-binding domain (Pfam accession: PF02362) and Auxin-response domain (Pfam accession: PF06507), were included in subsequent analyses (Additional file 2: Table S2).

Family phylogeny construction
To generate reliable sequence alignments for the collected ARF gene-family members, boundaries of the B3 and Auxin-response domains were identified by aligning each of the protein sequences onto the two HMM profiles using hmmpath v3.2.1 [47, 48]. Alignments of the two domains were separately refined using muscle v3.8.1551 [49] and concatenated to generate a broad-scale sequence alignment for ARF genes. Columns in the alignment with more than 20% gaps were removed using Phyutility v2.2.6 [50].

IQ-TREE v1.6.8 [51] software was employed to reconstruct the maximum likelihood (ML) gene tree. For the obtained broad-scale amino acid alignment, the JTT + R9 model was the best-fit evolutionary model selected by
the six angiosperm ARF genes were classified and labelled (clade I through VI) by comparing with those phylogenetically classified ARF genes. The phylogeny of angiosperm species and the associated paleopolyploidy events were redrawn based on a tree reported earlier by Van de Peer et al. [41] and the APG IV system [60] with minor modifications: the tetraploidy event in cucurbitaceae [61], the pentaploidy of the cotton genome [62], the fern genome duplications [45], the ancestral duplication events in mosses [4, 63] and in Caryophyllales [64], were included in the tree. The ARF syntenic networks were analyzed using CFinder v2.0.6 [38] utilizing the unweighted CPM algorithm and no time limit. All possible k-clique (from 3 to 21) communities were identified for the complete ARF gene syntenic network. We used k=3 as the clique community threshold and in this scenario one ARF gene (node) involved in a subnetwork community should have at least two connections (edge) with other nodes in the community. Increasing k values made the communities smaller and more disintegrated but also more connected. For illustration purposes, we used different nodal shapes to represent the members from the six ARF groups and different colors to depict specific plant lineages using the Cytoscape v3.7.0 software [58]. For each of the 25 communities, the species composition of the syntelogs were counted and a heatmap was generated using the pheatmap v1.0.10 (https://github.com/ravokolde/pheatmap) package implemented in the R statistical environment.

**Genomic synteny network construction**

To unveil the genomic syntenic relationships among plants, protein sequences for each of the 52 angiosperm genomes were compared with each other and themselves using Diamond v0.9.22.123 software [56] with an e-value cutoff at 1e–5. In this way, blastp tables for a total of $52 \times 51/2 + 52 = 2704$ whole proteome comparisons were generated. Only the top five non-self blastp hits were retained as input for the MCScanX [57] analyses. The ARF gene associated syntenic genomic blocks were extracted and visualized in Cytoscape v3.7.0 [58] and Gephi v0.9.2 [59]. Some ARF syntelogs were truncated or demonstrate absence of signature domains and were not included in our phylogenetic analyses. These truncated ARF genes were classified and labelled (clade I through VI) by comparing with those phylogenetically classified ARF genes. The phylogeny of angiosperm species and the associated paleopolyploidy events were redrawn based on a tree reported earlier by Van de Peer et al. [41] and the APG IV system [60] with minor modifications: the tetraploidy event in cucurbitaceae [61], the pentaploidy of the cotton genome [62], the fern genome duplications [45], the ancestral duplication events in mosses [4, 63] and in Caryophyllales [64], were included in the tree.

The ARF syntenic networks were analyzed using CFinder v2.0.6 [38] utilizing the unweighted CPM algorithm and no time limit. All possible k-clique (from 3 to 21) communities were identified for the complete ARF gene syntenic network. We used k=3 as the clique community threshold and in this scenario one ARF gene (node) involved in a subnetwork community should have at least two connections (edge) with other nodes in the community. Increasing k values made the communities smaller and more disintegrated but also more connected. For illustration purposes, we used different nodal shapes to represent the members from the six ARF groups and different colors to depict specific plant lineages using the Cytoscape v3.7.0 software [58]. For each of the 25 communities, the species composition of the syntelogs were counted and a heatmap was generated using the pheatmap v1.0.10 (https://github.com/ravokolde/pheatmap) package implemented in the R statistical environment.

**Additional file 1: Fig. S1.** Plant lineages screened for ARF homologues.

**Fig. S2.** Number of Auxin Response Factor genes identified from each of the plant genomes.

**Fig. S3.** Terminal branch length comparison between angiosperm and gymnosperm ARF genes.

**Fig. S4.** Phylogenetic and synteny network analyses for each of the six groups of ARF genes in angiosperms.

**Table S1.** Annotation and classification of ARF genes in Arabidopsis thaliana.

**Table S2.** List and classification of ARF genes analyzed in angiosperms.

**Abbreviations**

ARF: Auxin response factor; HMM: Hidden Markov Model; ML: Maximum likelihood; CPM: Critical path method; TMK: Transmembrane kinase.

**Acknowledgements**

We thank the School of Life Sciences of The Chinese University of Hong Kong for providing the computational resources.

**Authors’ contributions**

BG conceived the study, performed the bioinformatic analyses and wrote the manuscript. LW and MC contributed to the data collection and discussion. MO and JZ critically revised and improved the manuscript. All authors read and approved the final manuscript.

**Funding**

This work was supported by Hong Kong Research Grant Council (GRF 12100318, AoE/M-05/12 and AoE/M-403/16), the Fundamental Research Funds for the Central Non-profit Research Institution of CAF (No. CAFYB-B2018QB002), Shenzhen Institute of Synthetic Biology Scientific Research Program (ZTXM20190003) and the National Science Foundation (Collaborative Research: Dimensions: No. 1638972 to M.J.O.).

**Availability of data and materials**

The datasets supporting the conclusions of this article are included within the article and additional files.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.
Competing interests

The authors declare that they have no competing interests.

Author details

1 State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong, China. 2 State Key Laboratory of Tree Genetics and Breeding, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China. 3 USDA-ARS, Plant Genetics Research Unit, University of Missouri, Columbia, MO 65211, USA. 4 CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China. 5 Department of Biology, Faculty of Science, Hong Kong Baptist University, Hong Kong, China.

Received: 30 November 2019 Accepted: 30 April 2020 Published online: 14 May 2020

References

1. Kieffer M, Neve J, Kepinski S. Defining auxin response contexts in plant development. Curr Opin Plant Biol. 2010;13(1):12–20.
2. Finet C, Jalilis T. Auxology: when auxin meets plant evo-devo. Dev Biol. 2012;369(1):19–31.
3. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, et al. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science. 2008;319(5859):64–9.
4. Lang D, Ullrich KK, Murat F, Fuchs J, Jenkins J, Haas FB, Piednoel M, Gundlach H, Van Bel M, Meyberg R, et al. The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. Plant J Cell Mol Biol. 2018;93(3):515–33.
5. Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu S, Ishizaki K, Yamaoka S, Nishiyama R, Nakamura Y, Berger F, et al. Insights into land plant evolution garnered from the Marchantia polymorpha genome. Cell. 2017;171(2):287–304.e215.
6. Banks JA, Nishiyama T, Hasebe M, Bowman JL, Grisbok M, dePamphilis CW, Albert VA, Aono N, Aoyama T, Ambrose BA, et al. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science. 2011;332(6032):960–3.
7. Flores-Sandoval E, Eklund DM, Hong S-F, Alvarez JP, Fisher TJ, Lampugnani ER, Gole JF, Vázquez-Lobo A, Dierschke T, Lin S-S, et al. Class C ARFs evolved before the origin of land plants and antagonize differentiation and developmental transitions in Marchantia polymorpha. New Phytol. 2018;218(4):1612–30.
8. Mutte SK, Kato H, Rothfels C, Melkonian M, Wong GK, Weijers D. Origin and evolution of the auxin receptor system. Elife. 2018;7:e3399.
9. Martin-Arevalillo R, Thavenon E, Juge F, Vnos-Poyo T, Vernoux T, Parcy F, Dumas R. Evolution of the auxin response factors from charophyte ancestors. PLoS Genet. 2019;15(9):e1008400.
10. Thelander M, Landberg K, Sundberg E. Auxin-mediated developmental control in the moss Physcomitrella patens. J Exp Bot. 2018;69(2):277–90.
11. Morris JL, Puttick MN, Clark JW, Edwards D, Kennic P, Pressel S, Wellman CH, Yang Z, Schneider H, Donoghue PCJ. The timescale of early land plant evolution. Proc Natl Acad Sci USA. 2018;115(10):E2274–83.
12. Chapman EJ, Estelle M. Mechanism of auxin-regulated gene expression in plants. Annu Rev Genet. 2009;43:265–85.
13. Finet C, Berne-Dedieu A, Scutt CP, Marletaz F. Evolution of the ARF gene family in land plants: old domains, new tricks. Mol Biol Evol. 2013;30(1):45–56.
14. Guilloye TJ, Hagen G. Auxin response factors. Curr Opin Plant Biol. 2007;10(5):453–60.
15. Ulmasov T, Hagen G, Guilloye TJ. ARF1, a transcription factor that binds to auxin response elements. Science. 1997;276(5320):1865–8.
16. Ulmasov T, Hagen G, Guilloye TJ. Dimerization and DNA binding of auxin response factors. Plant J. 1999;19(3):109–19.
17. Vernoux T, Brunoud G, Faricot E, Morin V, Van den Daele H, Legrand J, Oliva M, Das P, Rainier A, Wells D, et al. The auxin signalling network translates dynamic input into robust patterning at the shoot apex. Mol Syst Biol. 2011;7:508.
18. Cao M, Chen R, Li P, Yu Y, Zheng R, Ge D, Zheng W, Wang X, Gu Y, Gelova Z, et al. TMK1-mediated auxin signalling regulates differential growth of the apical hook. Nature. 2019;568:240–3.
19. Perez-Rodriguez P, Riano-Pachon DM, Correa LG, Rensing SA, Kersten B, Mueller-Roeber B. PINTFDB: updated content and new features of the plant transcription factor database: Nucleic Acids Res. 2010;38(Database issue):D822–7.
20. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017;45(D1):D1040–5.
21. Zheng Y, Jiao C, Sun H, Rosi H, Pombo MA, Zhang P, Bard F, Dai X, Martin GB, Giovannoni JJ, et al. ITAK: a program for genome-wide prediction of plant transcription factors, transcriptional regulators, and protein kinases. Mol Plant. 2016;9(12):1667–70.
22. Willemsson PKI, Muhlich C, Ullrich KK, Rensing SA. Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. Genome Biol Evol. 2017;9(12):3384–97.
23. Finet C, Fourquin C, Vinauger M, Berne-Dedieu A, Chambrier P, Painda-voine S, Scutt CP. Parallel structural evolution of auxin response factors in the angiosperms. Plant J. 2010;63(3):952–9.
24. Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas FB, Vanderstraeten L, Becker D, Lang D, et al. The Choara genome: secondary complexity and implications for plant terrestrialization. Cell. 2018;174(2):448–464.e242.
25. Wang D, Pei K, Fu Y, Sun Z, Li S, Liu H, Tang K, Han B, Tao Y. Genome-wide analysis of the auxin response factors (ARF) gene family in rice (Oryza sativa). Gene. 2007;394(1–2):13–24.
26. Wang Y, Deng DX, Shi YT, Miao N, Bian YL, Yin ZT. Diversification, phyllogeny and evolution of auxin response factor (ARF) family: insights gained from analyzing maize ARF genes. Mol Biol Rep. 2012;39(3):2401–15.
27. Kalluri UC, Difazio SP, Brunner AM, Tuskan GA. Genome-wide analysis of Aux/IAA and ARF gene families in Populus trichocarpa. BMC Plant Biol. 2007;7:59.
28. Li FW, Melkonian M, Rothfels CJ, Villareal JC, Stevenson DW, Graham SW, Wong GKS, Pryer KM, Mathews S. Phytochrome diversity in green plants and the origin of canonical plant phytochromes. Nat Commun. 2015;6(1):7852.
29. Gao B, Chen MX, Li XS, Liang YQ, Zhu FY, Liu TY, Zhang DY, Wood AJ, Oliver MU, Zhang JH. Evolution by duplication: paleopolyploidy events in plants reconstructed by deciphering the evolutionary history of VZC transcription factors. BMC Plant Biol. 2018;18(1):256.
30. Cheng SF, van den Bergh E, Zeng P, Zhong X, Xu U, Liu X, Hofberger J, de Brujin S, Bhide AS, Kuelahoglu C, et al. The Tarenaya hassleriinae genome provides insight into reproductive trait and genome evolution of cruci-fer Plant Cell. 2013;25(8):2813–30.
31. Zhao T, Holmer R, de Brujin S, Angenent GC, van den Burg HA, Schranz ME. Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. Plant Cell. 2017;29(6):1278–92.
32. Zhao T, Schranz ME. Network approaches for plant phylogenomic syn-teny analysis. Curr Opin Plant Biol. 2017;36:129–34.
33. Leliaert F, Smith DR, Moreau H, Henon MD, Verbruggen H, Delwiche CF, De Clerck O. Phylogeny and molecular evolution of the green algae. Curr Rev Plant Sci. 2012;13(1):1–46.
34. Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci USA. 2014;111(45):E8589–65.
35. Smith SA, Donoghue MJ. Rates of molecular evolution are linked to life history in flowering plants. Science. 2008;322(5898):86–9.
36. Jiao Y, Wickett NJ, Ayyampalayam S, Chandler AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011;473(7345):97–100.
37. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. Science. 2008;320(5875):486–9.
38. Ardasevick P, Bolla G, Farkas U, Demesy L, Vicske T, FINDER: locating cliques and overlapping modules in biological networks. Bioinformatics. 2006;22(8):1021–3.
39. Gamboa-Tuz SD, Pereira-Santana A, Zhao T, Schranz ME, Castano E, Rodriguez-Zapata LC. New insights into the phylogeny of the TMBIM superfamily across the tree of life: comparative genomics and synteny networks reveal independent evolution of the BI and LFG families in plants. Mol Phylogenet Evol. 2018;126:266–78.
40. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Roff M, Ruzicka DR, Wafula E, Wickett NJ, et al. A genome tetraploidization associated with early diversification of the core eudicots. Genome Biol. 2012;13(1):R3.
41. Van de Peer Y, Milzarchi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet. 2017;18(7):411–24.
42. Gao B, Chen M, Li X, Zhang J. Ancient duplications and grass-specific transposition influenced the evolution of LEAFY transcription factor genes. Commun Biol. 2019;2(1):237.
43. Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux PM, Li FW, Melkonian B, Mavrodiev EV, Sun W, et al. 1KP: a phylodiverse genome sequencing plan. Gigascience. 2018;7(3):1–9.
44. Matarci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirrab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M, et al. Data access for the 1,000 plants (1KP) project. Gigascience. 2014;3(1):17.
45. Li FW, Brouwer P, Carretero-Paulet L, Cheng SF, de Vries J, Delaux PM, Ely A, Koppers N, Kuo LY, Li Z, et al. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nat Plants. 2018;4(7):460–72.
46. Marchler-Bauer A, Bo Y, Han LY, He JE, Lanczycki CJ, Lu SN, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017;45(D1):D200–3.
47. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol. 2011;7(10):e1002195.
48. Eddy SR. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy duplication events in the ancestor of Sphagnopsida (Bryophyta). New Phytol. 2016;209(3):1252–63.
49. Devos N, Szovenyi P, Weston DJ, Rothfels CJ, Johnson MG, Shaw AJ. Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnum (Bryophyta). Mol Biol Evol. 2018;35(1):16–26.
50. Wang J, Sun P, Li Y, Liu Y, Yang N, Yu J, Ma X, Sun S, Xia R, Liu X, et al. An overlooked paleotetraploidization in cucurbitaceae. Mol Biol Evol. 2016;33(1):16–26.
51. Wang X, Guo H, Wang J, Lei T, Liu Y, Wang Z, Li Y, Lee TH, Li J, Tang H, et al. Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation. New Phytol. 2016;209(3):1252–63.
52. Devos N, Szovenyi P, Weston DJ, Rothfels CJ, Johnson MG, Shaw AJ. Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnum (Bryophyta). Mol Biol Evol. 2018;35(1):16–26.
53. Kalyaanamoorthy S, Minh BQ, Wor J, von Haeseler A, Minh BQ. IQ-TREE: a fast, convenient online submission
42. Gao B, Chen M, Li X, Zhang J. Ancient duplications and grass-specific transposition influenced the evolution of LEAFY transcription factor genes. Commun Biol. 2019;2(1):237.
43. Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux PM, Li FW, Melkonian B, Mavrodiev EV, Sun W, et al. 1KP: a phylodiverse genome sequencing plan. Gigascience. 2018;7(3):1–9.
44. Matarci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirrab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M, et al. Data access for the 1,000 plants (1KP) project. Gigascience. 2014;3(1):17.
45. Li FW, Brouwer P, Carretero-Paulet L, Cheng SF, de Vries J, Delaux PM, Ely A, Koppers N, Kuo LY, Li Z, et al. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nat Plants. 2018;4(7):460–72.
46. Marchler-Bauer A, Bo Y, Han LY, He JE, Lanczycki CJ, Lu SN, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017;45(D1):D200–3.
47. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol. 2011;7(10):e1002195.
48. Eddy SR. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy duplication events in the ancestor of Sphagnopsida (Bryophyta). New Phytol. 2016;209(3):1252–63.
54. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44(W1):W242–5.
55. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
56. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(11):59–60.
57. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49.
58. Shannon R, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
59. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: International AAAI conference on weblogs and social media 2009. 2009: vol. 8, p. 361–2.
60. Byng JW, Chase MW, Christenhusz MJM, Fay MF, Judd WS, Mabberley DJ, Sennikov AN, Solits DE, Solits PS, Stevens PF, et al. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc. 2016;181(1):1–20.
61. Wang J, Sun P, Li Y, Liu Y, Yang N, Yu J, Ma X, Sun S, Xia R, Liu X, et al. An overlooked paleotetraploidization in cucurbitaceae. Mol Biol Evol. 2018;35(1):16–26.
62. Wang X, Guo H, Wang J, Lei T, Liu Y, Wang Z, Li Y, Lee TH, Li J, Tang H, et al. Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation. New Phytol. 2016;209(3):1252–63.
63. Devos N, Szovenyi P, Weston DJ, Rothfels CJ, Johnson MG, Shaw AJ. Analyses of transcriptome sequences reveal multiple ancient large-scale duplication events in the ancestor of Sphagnum (Bryophyta). Mol Biol Evol. 2018;35(1):16–26.
64. Yang Y, Moore MJ, Brockington SF, Mikenas J, Olivieri J, Walker JF, Smith SA. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. New Phytol. 2018;217(2):853–70.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.