One thousand plant transcriptomes and the phylogenomics of green plants

One thousand plant transcriptomes and the phylogenomics of green plants

https://doi.org/10.1038/s41586-019-1693-2

Received: 17 November 2017

Accepted: 12 September 2019

Published online: 23 October 2019

Open access

One Thousand Plant Transcriptomes Initiative

Green plants (Viridiplantae) include around 450,000–500,000 species[1,2] of great diversity and have important roles in terrestrial and aquatic ecosystems. Here, as part of the One Thousand Plant Transcriptomes Initiative, we sequenced the vegetative transcriptomes of 1,124 species that span the diversity of plants in a broad sense (Archaeplastida), including green plants (Viridiplantae), glaucophytes (Glaucophyta) and red algae (Rhodophyta). Our analysis provides a robust phylogenomic framework for examining the evolution of green plants. Most inferred species relationships are well supported across multiple species tree and supermatrix analyses, but discordance among plastid and nuclear gene trees at a few important nodes highlights the complexity of plant genome evolution, including polyploidy, periods of rapid speciation, and extinction. Incomplete sorting of ancestral variation, polyploidization and massive expansions of gene families punctuate the evolutionary history of green plants. Notably, we find that large expansions of gene families preceded the origins of green plants, land plants and vascular plants, whereas whole-genome duplications are inferred to have occurred repeatedly throughout the evolution of flowering plants and ferns. The increasing availability of high-quality plant genome sequences and advances in functional genomics are enabling research on genome evolution across the green tree of life.

Viridiplantae comprise an estimated 450,000–500,000 species[1,2], encompass a high level of diversity and evolutionary timescales[3], and have important roles in all terrestrial and most aquatic ecosystems. This ecological diversity derives from developmental, morphological and physiological innovations that enabled the colonization and exploitation of novel and emergent habitats. These innovations include multicellularity and the development of the plant cuticle, protected embryos, stomata, vascular tissue, roots, ovules and seeds, and flowers and fruit (Fig. 1). Thus, plant evolution ultimately influenced environments globally and created a cascade of diversity in other lineages that span the tree of life. Plant diversity has also fuelled agricultural innovations and growth in the human population[4].

Phylogenomic approaches are now widely used to resolve species relationships[5] as well as the evolution of genomes, gene families and gene function[6]. We used mostly vegetative transcriptomes for a broad taxonomic sampling of 1,124 species together with 31 published genomes to infer species relationships and characterize the relative timing of organismal, molecular and functional diversification across green plants.

We evaluated gene history discordance among single-copy genes. This is expected in the face of rapid species diversification, owing to incomplete sorting of ancestral variation between speciation events[7]. Hybridization[8], horizontal gene transfer[9], gene loss following gene and genome duplications[10] and estimation error can also contribute to gene-tree discordance. Nevertheless, through rigorous gene and species tree analyses, we derived robust species tree estimates (Fig. 2 and Supplementary Figs. 1–3). Gene-family expansions and genome duplications are recognized sources of variation for the evolution of gene function and biological innovations[11,12]. We inferred the timing of ancient genome duplications and large gene-family expansions. Our findings suggest that extensive gene-family expansions or genome duplications preceded the evolution of major innovations in the history of green plants.

Integrated analysis of genome evolution

Because genome sizes vary by 2,340-fold in land plants[13] and 4,680-fold in chlorophyte and streptophyte green algae[14], we used a reduced-representation sequencing approach to reconstruct gene and species histories. Specifically, we generated 1,342 transcriptomes representing 1,124 species across Archaeplastida, including green plants, glaucophytes and red algae. Comparing phylogenetic inferences based on nuclear and plastid genes (Figs. 2, 3 and Supplementary Figs. 1–3), we obtained well-supported, largely congruent results across diverse datasets and analyses. Resolution of some relationships, however, was confounded by gene-tree discordance (Fig. 3), which is attributable to factors that include rapid diversification, reticulate evolution, gene duplication and loss, and estimation error.

Inferred whole-genome duplications (WGDs; that is, polyploidy) across the gene-tree summary phylogeny estimated using ASTRAL[15] were not uniformly distributed (Fig. 4, Supplementary Fig. 8 and Supplementary Table 2). Comparing distributions of gene duplication times for each species[16] (Supplementary Table 3) and orthologue divergence times[17] (Supplementary Table 4) with gene-tree analyses[18] (Supplementary Tables 5, 6), we inferred 244 ancient WGDs across Viridiplantae (Supplementary Fig. 8 and Supplementary Table 2). Although there

A list of participants and their affiliations appears in the online version of the paper.

Fig. 1 | Diversity within the Viridiplantae. a–**e**, Green algae. **a**, *Acetabularia* sp. (Ulvophyceae). **b**, *Stephanosphaera pluvialis* (Chlorophyceae). **c**, *Botryococcus* sp. (Trebouxiophyceae). **d**, *Chara* sp. (Charophyceae). **e**, '*Spirotaenia*' sp. (taxonomy under review) (Zygnematophyceae). **f**–**p**, Land plants. **f**, *Notothylas orbicularis* (Anthocerotophyta (hornwort)). **g**, *Conocephalum conicum* (Marchantiophyta (thalloid liverwort)). **h**, *Sphagnum* sp. (Bryophyta (moss)). **i**, *Dendrolycopodium obscurum* (Lycopodiophyta (club moss)). **j**, *Equisetum telmateia* (Polypodiopsida, Equisetidae (horsetail)). **k**, *Parablechnum schiedeanum* (Polypodiopsida, Polypodiidae (leptosporangiate fern)). **l**, *Ginkgo biloba* (Ginkgophyta). **m**, *Pseudotsuga menziesii* (Pinophyta (conifer)). **n**, *Welwitschia mirabilis* (Gnetophyta). **o**, *Bulnesia arborea* (Angiospermae, eudicot, rosid). **p**, *Paphiopedilum lowii* (Angiospermae, monocot, orchid). **a**, Photograph reproduced with permission of Thieme Verlag, Stuttgart[66]. **b**–**e**, Photographs courtesy of M. Melkonian. **f**–**j**, **l**–**n**, **p**, Photographs courtesy of D.W.S. **k**, Photograph courtesy of R. Moran. **o**, Photograph courtesy of W. Judd.

are limitations to the inference of WGD events using this approach, we found that comparisons of these results with 65 overlapping published genome-based WGD inferences revealed 6 false-negative results in our tree-based estimates and no false-positive results (Supplementary Table 2). Analyses based on whole-genome sequences are needed for further resolution of WGD events.

With the exception of most *Selaginella* species and some liverworts (Fig. 1g), our analyses implicated at least one ancient WGD in the ancestry of every land plant lineage. By contrast, most algal lineages showed no evidence of WGD. Notably, the predicted sister clade of land plants (Fig. 2), Zygnematophyceae (Fig. 1e), exhibited the highest density of WGDs among algal lineages (Fig. 4), although the apparent increase in WGD was largely restricted to the desmid clade (Desmidiales) within Zygnematophyceae.

Increased diversification rates did not precisely co-occur with WGDs on the phylogeny. WGDs are expected to contribute to the evolution of novel gene function[11,12]. For example, novel functions among duplicate MADS-box genes that arose through WGD have been linked to the origin of flowering plants[19,20] and core eudicots[21], and functional diversification of gene families after WGD has contributed to the evolution of fruit colour in tomato species[22,23] and to nodule development within legumes[22,24]. Consistent with previous studies with less extensive taxon sampling[24–27], however, we inferred lags between WGDs and increased species diversity. Integrated phylogenomic and functional investigations are required to gain a mechanistic understanding of the lag between WGD, the evolution of novel gene functions and their potential influence on diversification rates.

Gene-family expansions (and contractions) contribute to the dynamic evolution of metabolic, regulatory and signalling networks[28,29]. Given the inherent limitations of transcriptome data, we searched for large-fold changes in 23 of the largest gene families in *Arabidopsis thaliana*[30] that are involved in many important functions (such as transcriptional regulation, enzymatic and signalling function, and transport; Fig. 5 and Supplementary Tables 7, 8). Although our RNA-sequencing-based sampling of expressed genes is incomplete, the median representation of universally conserved genes[31] was 80–90% for taxa across Viridiplantae (Extended Data Fig. 3a, b). Furthermore, there was a strong correlation ($r = 0.95$) between gene-family sizes in our transcriptomes (focusing on the largest gene families) and those of fully sequenced genomes (Extended Data Fig. 3c–f). We identified gene-family expansions and contractions, including some that have been described previously[32–34]. Specifically, the *AP2*, *bHLH*, *bZip* and *WRKY* transcription factor families were inferred to be present in the last common ancestor of Viridiplantae, whereas the origin of *GRAS* and *NAC* genes occurred in early streptophytes after divergence from the chlorophyte algal lineage (Fig. 5). The highest concentration of expansion events was inferred along the 'spine' of the phylogeny between the origins of Viridiplantae and vascular plants (Fig. 5b and Supplementary Table 7). Expansions of some focal gene families also continued after the origin of embryophytes; however, no expansions occurred in association with the origin and radiation of angiosperms (Fig. 5). Gene-family expansions and functional diversification may have contributed to the adaptations required for life in terrestrial habitats, but the sizes of these focal gene families apparently stabilized in the face of continued gene duplication and loss throughout the evolution of vascular plants.

Primary acquisition of the plastid

The primary acquisition of the plastid in an ancestor of extant Archaeplastida was a pivotal event in the history of life. All possible relationships among Viridiplantae, Glaucophyta and Rhodophyta have been hypothesized, with alternative implications for the gain and loss of characters[35] in the early history of the three lineages. Strong support for the sister relationship of Viridiplantae and Glaucophyta[35] (Figs. 2, 3a) found here indicates that ancestral red algae lost flagella and peptidoglycan biosynthesis, perhaps associated with a reduction in genome size[36]. Peptidoglycan biosynthesis was independently lost early in the evolution of Chlorophyta[37] and within angiosperms[38].

The history of Viridiplantae

The origin of Viridiplantae is marked by the loss of light-harvesting phycobilisomes composed of phycobiliproteins, the evolution of the accessory photosynthetic pigment chlorophyll *b*, which has a distinct light-absorption spectrum relative to chlorophyll *a*, and intraplastidial starch synthesis and deposition. Viridiplantae are consistently recovered as monophyletic, with early diverging Chlorophyta and Streptophyta lineages[39–41]. However, the placement of the picoplanktonic algal lineage Prasinococcales was unstable in our analyses (Fig. 3e).

Diversification within Chlorophyta

All nuclear-gene analyses resolved a grade of largely marine unicellular lineages subtending the core clade consisting of Trebouxiophyceae, Ulvophyceae and Chlorophyceae[42] (Fig. 1a–c and Supplementary Figs. 1–3). The nuclear supermatrix and ASTRAL trees placed Trebouxiophyceae as sister to a clade containing Chlorophyceae and Ulvophyceae[42,43]. However, whereas the supermatrix trees supported Ulvophyceae as monophyletic, the ASTRAL tree resolved Ulvophyceae as a grade and Bryopsidales is poorly supported as sister to Chlorophyceae (Fig. 3h). All tree estimates suggest that there were multiple origins of multicellularity

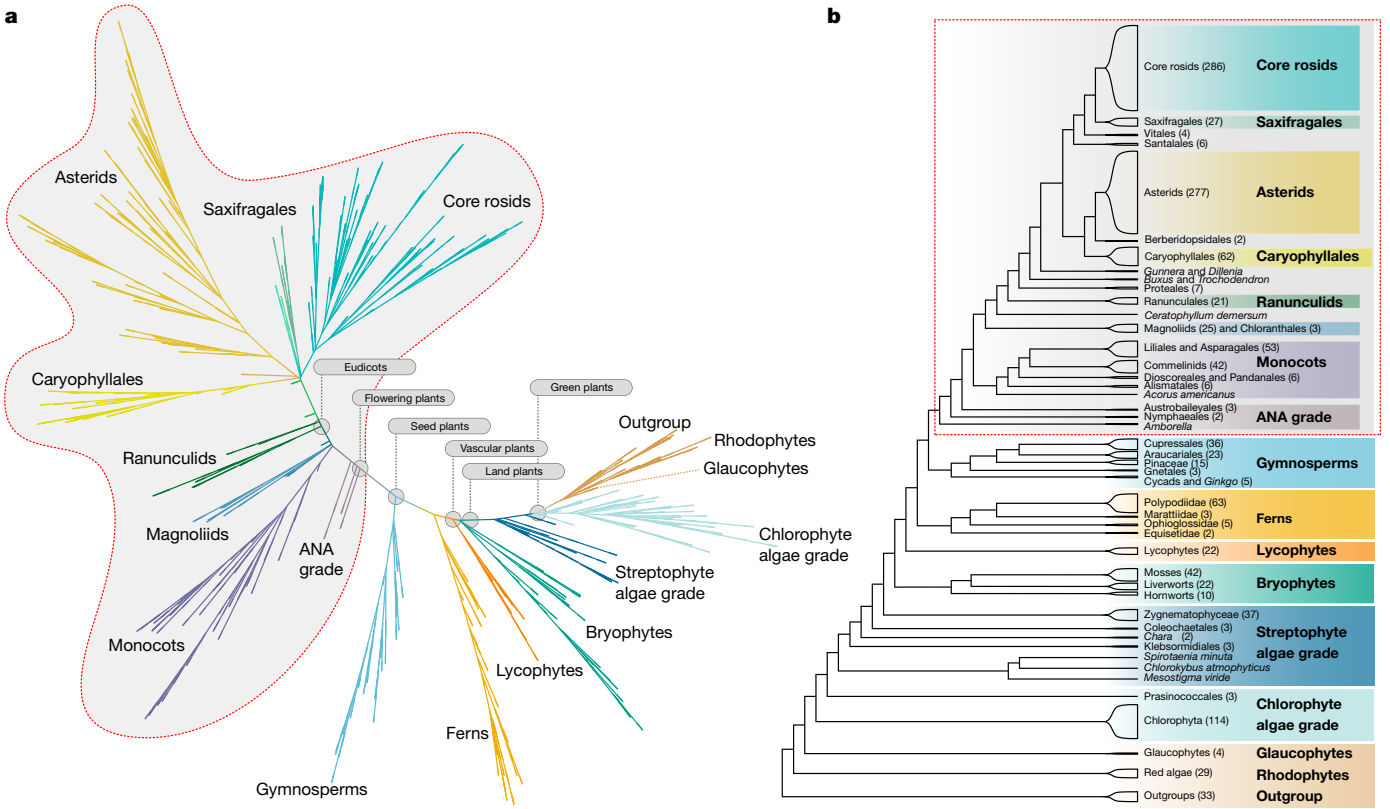


Fig. 2 | Phylogenetic inferences of major clades. Phylogenetic inferences were based on ASTRAL analysis of 410 single-copy nuclear gene families extracted from genome and transcriptome data from 1,153 species, including 1,090 green plant (Viridiplantae) species (Supplementary Table 1). **a**, Phylogram showing internal branch lengths proportional to coalescent units ($2N_e$ generations) between branching events, as estimated by ASTRAL-II[15] v.5.0.3. **b**, Relationships among major clades with red box outlining flowering plant clade. Species numbers are shown for each lineage. Most inferred relationships were robust across data types and analyses (Supplementary Figs. 1–3) with some exceptions (Supplementary Fig. 6). Data and analysis scripts are available at https://doi.org/10.5281/zenodo.3255100.

within Ulvophyceae. Only 12 out of 119 sampled chlorophyte species exhibited evidence of a WGD in their ancestry, and most of these putative WGDs were restricted to single clades.

Streptophyta

The evolution of streptophytes was associated with several adaptations to terrestrial habitats[44–46]. All analyses recovered *Mesostigma*, *Chlorokybus* and *Spirotaenia minuta* in a clade that is sister to the remainder of Streptophyta[39] with successive divergence of Klebsormidiales, Charophyceae (Fig. 1d), Coleochaetophyceae and Zygnematophyceae (Fig. 1e) relative to Embryophyta. However, with greatly increased taxon sampling relative to our previous work[39], internal branch lengths are diminished, and we could not reject the possibility of a true radiation giving rise to Coleochaetales, Zygnematophyceae and embryophyte lineages (land plants; Figs. 1f–p, 3g(II)). Although quartet support for a clade of Coleochaetales and Zygnematophyceae as sister to embryophytes was similar to support for Zygnematophyceae as sister to embryophytes, a clade consisting of Coleochaetales and land plants was not supported.

Embryophyta

Land plants include many of the most familiar green plants (for example, bryophytes (Fig. 1f–h), lycophytes (Fig. 1i), ferns (Fig. 1j, k) and seed plants (Fig. 1l–p)). They exhibit key innovations, including protected reproductive organs (archegonia and antheridia) and the development of the zygote within an archegonium into an embryo that receives maternal nutrition. Resolving relationships among bryophytes (mosses, liverworts and hornworts) and their relationships to the remaining land plants has long been problematic, but is critical for understanding the evolution of fundamental innovations within land plants, including the tolerance to desiccation, shifts in the dominance of multicellular haploid and diploid generations, and parental retention of a multicellular embryo.

Bryophytes have sometimes been resolved as a grade[47,48], with liverworts, mosses and hornworts as successive sister groups to Tracheophyta (vascular plants; Fig. 1i–p). We recovered extant bryophytes as monophyletic in the ASTRAL analysis of nuclear gene trees (Fig. 3b) and plastome analyses, with hornworts sister to a moss and liverwort clade. All analyses rejected the hypothesis that liverworts are sister to all other extant land plant lineages[39,49].

The largest number of gene-family expansions in our analyses was associated with the origin of land plants and the evolution of bryophytes (transition between streptophyte algae and bryophytes in Fig. 5b). By contrast, we found no evidence of WGD on the stem branch for land plants (Supplementary Tables 5, 6).

Vascular plants

Within the vascular plants, lycophytes are supported as the sister group of Euphyllophyta (ferns and seed plants). We found no evidence of pan-vascular-plant or ancestral euphyllophyte WGDs, but some gene-family expansions were associated with the origin of vascular plants (Fig. 5b).

Within ferns (Polypodiopsida), plastid data weakly support Equisetales as sister to Psilotales and Ophioglossales (Supplementary Fig. 3), whereas nuclear gene analyses robustly place Equisetales sister to the remaining ferns[50]. The supermatrix and plastome-based trees placed Marattiales sister to the leptosporangiate ferns[50] (Polypodiidae), but ASTRAL recovers nearly equal quartet support for this hypothesis or

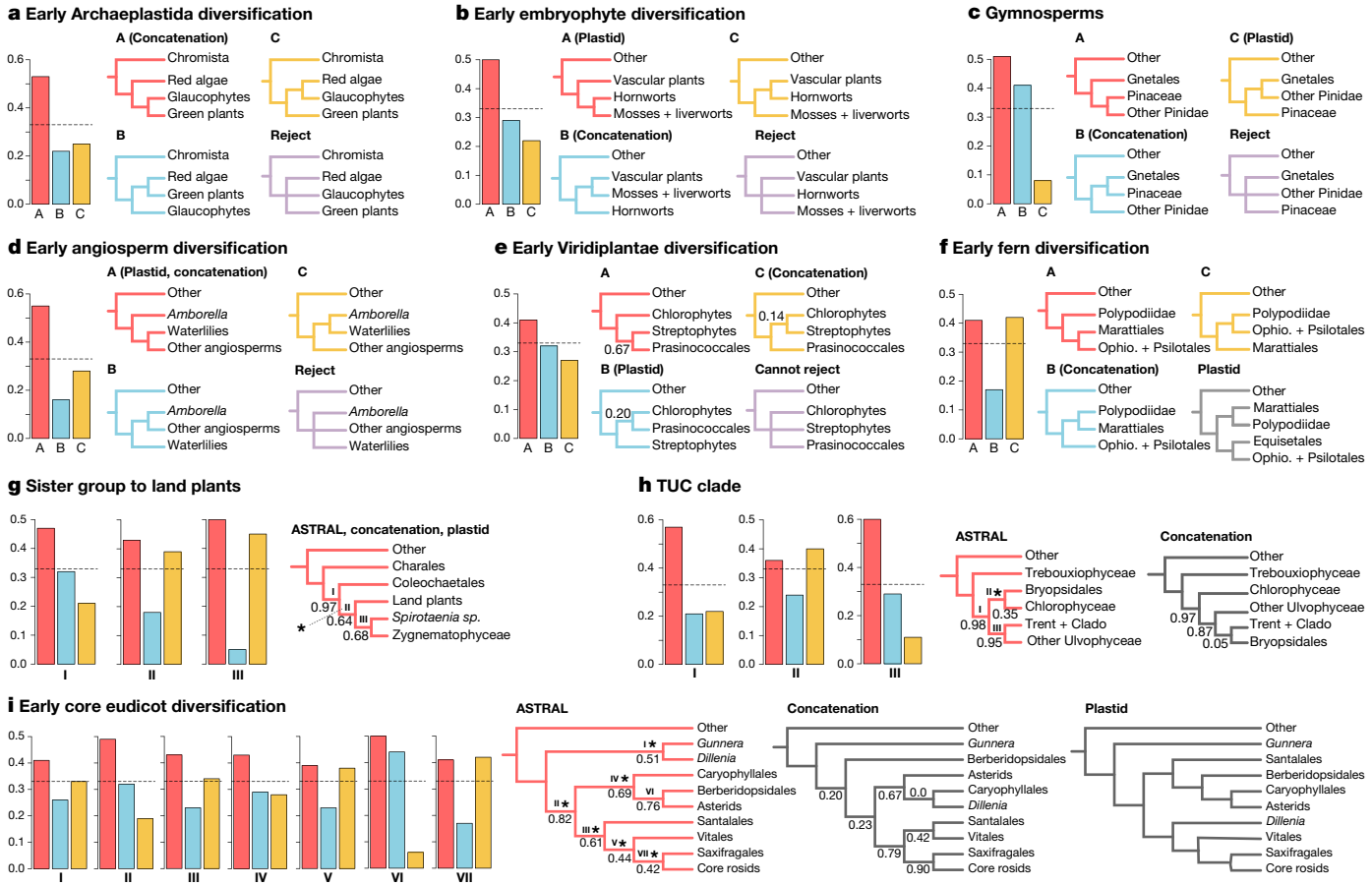


Fig. 3 | Alternative branching orders for contentious relationships. Local posterior probabilities (shown only when below 1.0) and gene-tree quartet frequencies (bar graphs) for alternative branching orders for contentious relationships in the plant phylogeny (see text). **a**, Early Archaeplastida diversification. **b**, Early embryophyte diversification. **c**, Gymnosperms. **d**, Early angiosperm diversification. **e**, Early Viridiplantae diversification. **f**, Early fern diversification. **g**, The sister lineage to land plants. **h**, Trebouxiophyceae, Ulvophyceae and Chlorophyceae. **i**, Eudicot diversification. Red bars represent the ASTRAL topology; blue and yellow trees and bars represent the frequencies of alternative branching orders in ASTRAL. The topologies recovered in the concatenated supermatrix analysis and plastid gene analyses are also indicated. Dashed horizontal lines mark expectation for a hard polytomy (purple). In **g**–**i**, panels include more than 4 tips, so nodes are delineated with Roman numerals and bar graphs are shown for each node and asterisks above branches indicate failure to reject the hypothesis that the node is a polytomy. Data and analysis scripts are available at https://doi.org/10.5281/zenodo.3255100.

for Marattiales as sister to Psilotales and Ophioglossales (Fig. 3f). Leptosporangiate ferns (Fig. 1k) experienced more WGD events than any other lineage of Viridiplantae outside the angiosperms, with an average of 3.79 inferred WGDs in the history of each sampled species (Fig. 4). WGD was inferred in an ancestor of all extant ferns and an additional 19 putative WGDs were implicated in the ancestry of fern subclades (Ophioglossaceae and Polypodiaceae; Fig. 4, Supplementary Fig. 8 and Supplementary Tables 2, 5, 6). Considering the high chromosome numbers of some ferns, our discovery that they exhibit one of the highest frequencies of palaeopolyploidization among green plants is not unexpected[51].

Whereas none of our focal gene families exhibited significant expansion in ferns, significantly more MIKC-type MADS-box genes—involved in specification of ovule and flower development in seed plants[52]—were observed in leptosporangiate ferns relative to all other green plant lineages, other than seed plants (Extended Data Fig. 1). The ancestral number of MIKC-type MADS-box genes for ferns and seed plants was 4 or 5, and gene numbers increased independently within leptosporangiate ferns and seed plants (Extended Data Figs. 1, 2).

Seed plants

A WGD in the ancestry of all extant seed plants has been inferred previously[18,53] but remains contested[54]. Gene-tree[18] analyses revealed significantly more gene duplications on the branch leading to extant seed plants than expected from background gene birth and death rates (analyses D1 ($P < 2.0 \times 10^{-18}$) and D2 ($P < 8.9 \times 10^{-16}$) in Supplementary Table 5). Numerous gene-family expansions were also associated with the origin of seed plants, and only one contraction was detected among the gene families analysed (Fig. 5b). Type II MIKC-type MADS-box genes exhibited a nearly twofold expansion independent of their expansion in ferns (Extended Data Figs. 1, 2).

Extant gymnosperms (approximately 1,000 species) are sister to flowering plants, and all of our analyses recovered Cycadales and *Ginkgo* (Fig. 1l) as a sister clade to the remaining gymnosperms (Fig. 3c). The placement of Gnetales conflicts strongly among the ASTRAL, supermatrix and plastome-based trees. Plastid data strongly support the 'Gnecup' hypothesis, with Gnetales as sister to a clade comprising Araucariales and Cupressales[47], whereas the supermatrix analysis of nuclear genes supports a 'Gnepine' hypothesis with Gnetales as sister to Pinales[55,56]. ASTRAL analyses strongly support the 'Gnetifer' hypothesis, with conifers (Araucariales, Cupressales and Pinales) sister to Gnetales[57]. The short internal branches in the ASTRAL tree suggest rapid diversification (Fig. 2). However, the uneven frequencies of gene-tree quartets—which support the alternative Gnecup and Gnepine hypotheses—suggest that gene-tree estimation biases[58] associated with increased substitution rates in Gnetales[59] or gene flow are possible sources of gene-tree

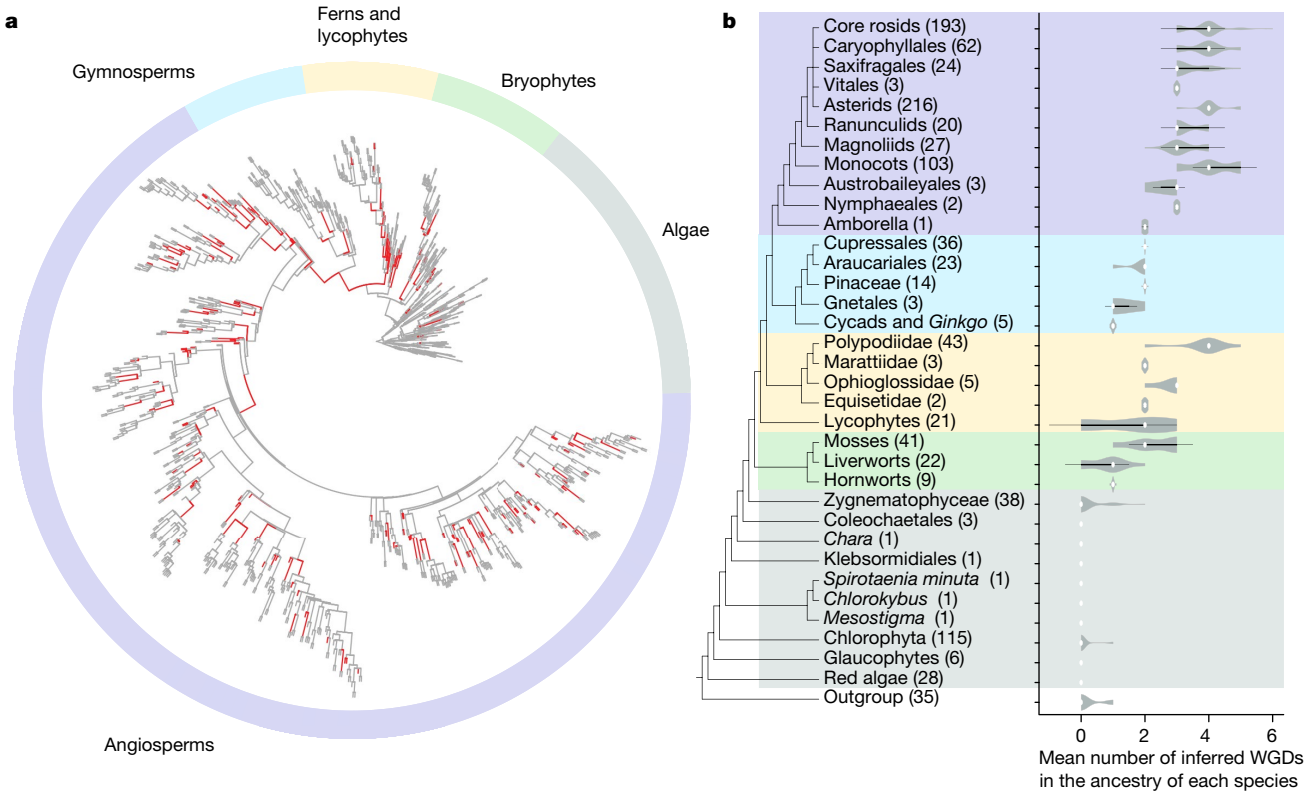


Fig. 4 | The distribution of inferred ancient WGDs across lineages of green plants. a, The locations of estimated WGDs are labelled red in the phylogeny of all 1000 Plants (1KP) samples. **b**, The number of inferred ancient polyploidization events within each lineage is shown in the violin plots. The white dot indicates the median, the thick black bars represent the interquartile range, the thin black lines define the 95% confidence interval and the grey shading represents the density of data points. The sample sizes for each lineage are shown within parentheses along with taxon names on the phylogeny. The phylogenetic placement of inferred WGDs is illustrated in Supplementary Fig. 8 and data supporting each WGD inference are provided in Supplementary Table 2.

discordance[8]. Previously inferred WGDs in ancestors of *Welwitschia*, Pinaceae and Cupressales[18] are supported, as is a new inference of WGD in the ancestry of Podocarpaceae (Fig. 4 and Supplementary Tables 2, 5, 6).

Angiosperms are by far the largest clade of green plants (more than 370,000 species[2]) and are marked by multiple key innovations, including the carpel, double fertilization, endosperm, and for most angiosperms, vessel elements. Both nuclear and plastid phylogenomic analyses agree with previous studies[39] in providing strong support for angiosperm monophyly and in placements of Amborellales, Nymphaeales and Austrobaileyales as successive sisters to all other angiosperms (Figs. 2, 3). Chloranthales and magnoliids comprise a clade in the ASTRAL and supermatrix analyses, but were resolved with poor support as successive sister lineages to all other Mesangiospermae (monocots, *Ceratophyllum* and eudicots) in the plastome-based tree. Whereas *Ceratophyllum* is sister to eudicots in the ASTRAL and plastome trees, it is poorly supported as sister to monocots in the supermatrix tree (Supplementary Figs. 1–3). All analyses suggest short time intervals between branching of the monocots, Magnoliidae, Chloranthales, Ceratophyllales and eudicot lineages in early mesangiosperm history (Fig. 2 and Supplementary Figs. 1–3).

Pentapetalae (70% of all angiosperms) are marked by the evolution of the pentamerous flower. Substantial gene-tree discordance was observed for relationships among core rosids, Saxifragales, Vitales, *Dillenia*, Santalales, Berberidopsidales, Caryophyllales, asterids and Gunnerales (the sister group of Pentapetalae; Fig. 3i). Short internal branches and poor support in the ASTRAL tree at the base of the core eudicots (Figs. 2, 3i) indicate rapid diversification following two rounds of WGD that resulted in palaeohexaploidy preceding the origin of the clade[60,61] (Supplementary Fig. 8). The supermatrix and plastid trees conflict with the poorly supported ASTRAL branching order (Fig. 3i). With the exception of the Berberidopsidales and core asterid clade, we were not able to reject the possibility of polytomies at the evaluated nodes in ASTRAL analyses (Fig. 3i).

Genomic and phylogenomic analyses have identified numerous WGDs throughout angiosperm history[62,63]. We found evidence that extant flowering plants descend from a polyploid common ancestor[19,53]. Gene-tree analyses detected a significantly larger-than-background number of gene duplications on the branch leading to the last common ancestor of extant angiosperms after divergence from the extant gymnosperm clade (analyses E1 ($P < 1.8 \times 10^{-41}$) and E2 (1.4×10^{-24}) in Supplementary Table 5). Furthermore, the numbers of inferred duplications on the stem branch of angiosperms were consistent with expectations for WGD (analyses E1 and E2 in Supplementary Table 6). We inferred over 180 WGDs within flowering plants, including 132 in eudicots and 35 in monocots (Supplementary Table 2).

The origin of the angiosperms was preceded by three focal gene-family contractions and no expansions (Fig. 5b), consistent with the hypothesis that the innovations in angiosperms may have involved the functional co-option of genes that were duplicated earlier in the evolution of seed plants[19]. We find that orthologues of some floral homeotic MADS-box genes originated in the stem group of extant seed plants approximately 300 million years ago (Extended Data Fig. 2), supporting the hypothesis that the origin of the angiosperm flower involved recruitment of developmental regulators that already existed in their seed plant ancestors[19,64].

Synthesis

These analyses establish a foundation for advancing our understanding of the overall phylogenetic framework of green plants and the genetic changes that were responsible for the characteristic traits associated with major evolutionary transitions in Viridiplantae. Portions of the

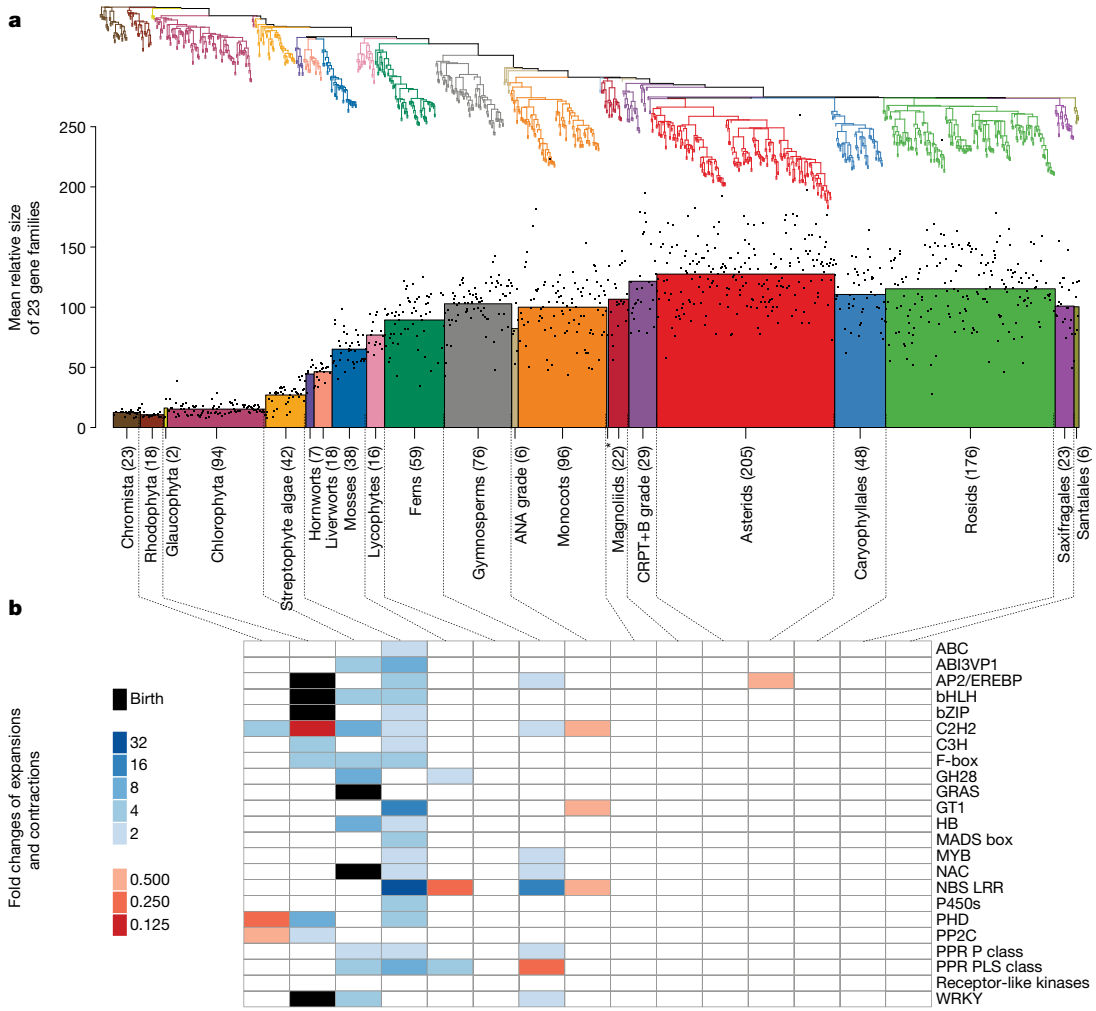


Fig. 5 | Assessment of significant expansions and contractions of largest plant gene families. a, Weighted average gene-family size for species groups (normalized to account for differences in gene-family sizes, weight = 1/(maximum observed gene-family size)). The ANA grade comprises Amborellales, Nymphaeales and Austrobaileyales, successive sister lineages to a clade with the remaining extant angiosperms; the 'CRPT+B' grade includes Ceratophyllales, Ranunculales, Proteales lineages and a Trochodendrales + Buxales clade in the ASTRAL tree (Fig. 2). Sample sizes are proportional to bar widths (from left to right, $n = 23$ (Chromista), 18 (Rhodophyta), 2 (Glaucophyta), 94 (Chlorophyta), 42 (streptophyte algae), 7 (hornworts), 18 (liverworts), 38 (mosses), 16 (lycophytes), 59 (ferns; monilophytes), 76(gymnosperms), 6 (ANA grade), 96 (monocots), 1 (*representing Chloranthales), 22 (magnoliids), 29 (CRPT+B grade), 205 (asterids), 48 (caryophyllids), 176 (rosids), 23 (Saxifragales) and 6 (Santalales). **b**, Gene families exhibiting significant copy number changes (two-sided Kolmogorov–Smirnov test; $P < 1 \times 10^{-6}$; gene-family expansions represent a gain of more than 50% and contractions represent a loss of more than 33%) with colour codes showing the magnitude of the observed fold changes. Data and analysis scripts are available at https://github.com/GrosseLab/OneKP-gene-family-evo.

species tree reported here remain unresolved. Phylogenetic analyses of genes extracted from a broad sampling of whole-genome sequences may improve gene family circumscriptions and resolve the species tree further. Expanded genome sequencing may also help to accurately account for interspecific gene flow, and orthology in the face of gene duplications and losses. However, for some nodes in the species tree, extensive discordance among inferred gene histories suggests that rapid diversification may not always conform to strict bifurcation of ancestral species into two descendent species.

Gene and genome duplications have long been considered a source of evolutionary novelty[11,12], producing an expanded molecular repertoire for adaptive evolution of key pathways and shifts in plant development and ecology. However, the direct connections between key innovations and specific gene duplications are rarely known, due in part to lag times between duplications and such inovations[25–27]. Phylogenetically informed experimental investigations of changes in gene content and function will improve our understanding of the roles of gene and genome duplications in the evolution of key innovations. Such efforts are underway, drawing on an expanding number of experimental model species distributed across the green plant tree of life[65].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1693-2

1. Corlett, R. T. Plant diversity in a changing world: status, trends, and conservation needs. *Plant Divers.* **38**, 10–16 (2016).
2. Lughadha, E. N. et al. Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. *Phytotaxa* **272**, 82–88 (2016).
3. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
4. Schery, R. W. *Plants for Man* 2nd edn (Prentice-Hall, 1972).
5. Philippe, H., Delsuc, F., Brinkmann, H. & Lartillot, N. Phylogenomics. *Annu. Rev. Ecol. Evol. Syst.* **36**, 541–562 (2005).

6. Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8**, 163–167 (1998).
7. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009).
8. Solís-Lemus, C., Yang, M. & Ané, C. Inconsistency of species tree methods under gene flow. *Syst. Biol.* **65**, 843–851 (2016).
9. Yang, Z. et al. Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation. *Proc. Natl Acad. Sci. USA* **113**, E7010–E7019 (2016).
10. Rasmussen, M. D. & Kellis, M. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* **22**, 755–765 (2012).
11. Ohno, S. *Evolution by Gene Duplication* (Springer-Verlag, 1970).
12. Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
13. Leitch, I. J. & Leitch, A. R. in *Plant Genome Diversity* Vol. 2 (eds Greilhuber, J. et al.) 307–322 (Springer, 2013).
14. Kapraun, D. F. Nuclear DNA content estimates in green algal lineages: chlorophyta and streptophyta. *Ann. Bot.* **99**, 677–701 (2007).
15. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015).
16. Barker, M. S. et al. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* **25**, 2445–2455 (2008).
17. Barker, M. S. et al. EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics. *Evol. Bioinform. Online* **6**, 143–149 (2010).
18. Li, Z. et al. Early genome duplications in conifers and other seed plants. *Sci. Adv.* **1**, e1501084 (2015).
19. Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
20. Ruelens, P. et al. The origin of floral organ identity quartets. *Plant Cell* **29**, 229–242 (2017).
21. Vekemans, D. et al. Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Mol. Biol. Evol.* **29**, 3793–3806 (2012).
22. Vanneste, K., Maere, S. & Van de Peer, Y. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Phil. Trans. R. Soc.B* **369**, 20130353 (2014).
23. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
24. Cannon, S. B. et al. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Mol. Biol. Evol.* **32**, 193–210 (2015).
25. Schranz, M. E., Mohammadin, S. & Edger, P. P. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Curr. Opin. Plant Biol.* **15**, 147–153 (2012).
26. Tank, D. C. et al. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol.* **207**, 454–467 (2015).
27. Landis, J. B. et al. Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* **105**, 348–363 (2018).
28. Maere, S. et al. Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102**, 5454–5459 (2005).
29. Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K. & Shiu, S.-H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* **148**, 993–1003 (2008).
30. Nelson, D. & Werck-Reichhart, D. A P450-centric view of plant evolution. *Plant J.* **66**, 194–211 (2011).
31. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
32. Bowman, J. L. et al. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**, 287–304 (2017).
33. Catarino, B., Hetherington, A. J., Emms, D. M., Kelly, S. & Dolan, L. The stepwise increase in the number of transcription factor families in the Precambrian predated the diversification of plants on land. *Mol. Biol. Evol.* **33**, 2815–2819 (2016).
34. Wilhelmsson, P. K. I., Mühlich, C., Ullrich, K. K. & Rensing, S. A. Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. *Genome Biol. Evol.* **9**, 3384–3397 (2017).
35. Rodríguez-Ezpeleta, N. et al. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr. Biol.* **15**, 1325–1330 (2005).
36. Qiu, H., Price, D. C., Yang, E. C., Yoon, H. S. & Bhattacharya, D. Evidence of ancient genome reduction in red algae (Rhodophyta). *J. Phycol.* **51**, 624–636 (2015).
37. van Baren, M. J. et al. Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. *BMC Genomics* **17**, 267 (2016).
38. Grosche, C. & Rensing, S. A. Three rings for the evolution of plastid shape: a tale of land plant FtsZ. *Protoplasma* **254**, 1879–1885 (2017).
39. Wickett, N. J. et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl Acad. Sci. USA* **111**, E4859–E4868 (2014).
40. Lewis, L. A. & McCourt, R. M. Green algae and the origin of land plants. *Am. J. Bot.* **91**, 1535–1556 (2004).
41. Becker, B. & Marin, B. Streptophyte algae and the origin of embryophytes. *Ann. Bot.* **103**, 999–1004 (2009).
42. Marin, B. Nested in the Chlorellales or independent class? Phylogeny and classification of the Pedinophyceae (Viridiplantae) revealed by molecular phylogenetic analyses of complete nuclear and plastid-encoded rRNA operons. *Protist* **163**, 778–805 (2012).
43. Cocquyt, E., Verbruggen, H., Leliaert, F. & De Clerck, O. Evolution and cytological diversification of the green seaweeds (Ulvophyceae). *Mol. Biol. Evol.* **27**, 2052–2061 (2010).
44. Delaux, P.-M. et al. Algal ancestor of land plants was preadapted for symbiosis. *Proc. Natl Acad. Sci. USA* **112**, 13390–13395 (2015).
45. Maugarny-Calès, A. et al. Apparition of the NAC transcription factors predates the emergence of land plants. *Mol. Plant* **9**, 1345–1348 (2016).
46. Delwiche, C. F. & Cooper, E. D. The evolutionary origin of a terrestrial flora. *Curr. Biol.* **25**, R899–R910 (2015).
47. Nickrent, D. L., Parkinson, C. L., Palmer, J. D. & Duff, R. J. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* **17**, 1885–1895 (2000).
48. Shaw, A. J., Szövényi, P. & Shaw, B. Bryophyte diversity and evolution: windows into the early evolution of land plants. *Am. J. Bot.* **98**, 352–369 (2011).
49. Puttick, M. N. et al. The interrelationships of land plants and the nature of the ancestral embryophyte. *Curr. Biol.* **28**, 733–745 (2018).
50. Rothfels, C. J. et al. The evolutionary history of ferns inferred from 25 low-copy nuclear genes. *Am. J. Bot.* **102**, 1089–1107 (2015).
51. Barker, M. S. & Wolf, P. G. Unfurling fern biology in the genomics age. *Bioscience* **60**, 177–185 (2010).
52. Theißen, G. & Gramzow, L. in *Plant Transcription Factors: Evolutionary, Structural, and Functional Aspects* (ed. Gonzalez, D. H.) 127–138 (Academic, 2016).
53. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
54. Ruprecht, C. et al. Revisiting ancestral polyploidy in plants. *Sci. Adv.* **3**, e1603195 (2017).
55. Bowe, L. M., Coat, G. & dePamphilis, C. W. Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proc. Natl Acad. Sci. USA* **97**, 4092–4097 (2000).
56. Chaw, S. M., Parkinson, C. L., Cheng, Y., Vincent, T. M. & Palmer, J. D. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl Acad. Sci. USA* **97**, 4086–4091 (2000).
57. Chaw, S. M., Zharkikh, A., Sung, H. M., Lau, T. C. & Li, W. H. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. *Mol. Biol. Evol.* **14**, 56–68 (1997).
58. Zhong, B., Yonezawa, T., Zhong, Y. & Hasegawa, M. The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol. Biol. Evol.* **27**, 2855–2863 (2010).
59. Wan, T. et al. A genome for gnetophytes and early evolution of seed plants. *Nat. Plants* **4**, 82–89 (2018).
60. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
61. Jiao, Y. et al. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3 (2012).
62. Soltis, D. E. et al. Polyploidy and angiosperm diversification. *Am. J. Bot.* **96**, 336–348 (2009).
63. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
64. Moyroud, E. et al. A link between LEAFY and B-gene homologues in *Welwitschia mirabilis* sheds light on ancestral mechanisms prefiguring floral development. *New Phytol.* **216**, 469–481 (2017).
65. Chang, C., Bowman, J. L. & Meyerowitz, E. M. Field guide to plant model systems. *Cell* **167**, 325–339 (2016).
66. Berger, S. & Kaever, M. J. *Dasycladales: an Illustrated Monograph of a Fascinating Algal Order* (Thieme Verlag, 1992).

One Thousand Plant Transcriptomes Initiative

James H. Leebens-Mack[1,135]*, Michael S. Barker[2,135], Eric J. Carpenter[3,134], Michael K. Deyholos[4,135], Matthew A. Gitzendanner[5,6,134], Sean W. Graham[7,135], Ivo Grosse[8,11,135], Zheng Li[2,134], Michael Melkonian[9,135], Siavash Mirarab[10,134,135], Martin Porsch[11,134], Marcel Quint[12,135], Stefan A. Rensing[13,14,135], Douglas E. Soltis[6,15,135], Pamela S. Soltis[6,15,135], Dennis W. Stevenson[16,135], Kristian K. Ullrich[17,134], Norman J. Wickett[18,19], Lisa DeGironimo[16,134], Patrick P. Edger[20,134], Ingrid E. Jordon-Thaden[5,6,21,134], Steve Joya[7,134], Tao Liu[22,134], Barbara Melkonian[9,134], Nicholas W. Miles[23,134], Lisa Pokorny[24,25,26,134], Charlotte Quigley[27,134], Philip Thomas[28,134], Juan Carlos Villarreal[29,134], Megan M. Augustin[30], Matthew D. Barrett[31,32,33], Regina S. Baucom[34], David J. Beerling[35], Ruben Maximilian Benstein[36], Ed Biffin[37], Samuel F. Brockington[38], Dylan O. Burge[39], Jason N. Burris[40,41,42], Kellie P. Burris[40,43], Valérie Burtet-Sarramegna[44], Ana L. Caicedo[45], Steven B. Cannon[46], Zehra Çebi[9], Ying Chang[7,47], Caspar Chater[48], John M. Cheeseman[49], Tao Chen[50], Neil D. Clarke[51], Harmony Clayton[52], Sarah Covshoff[38], Barbara J. Crandall-Stotler[53], Hugh Cross[54], Claude W. dePamphilis[55,134], Joshua P. Der[56], Ron Determann[57], Rowan C. Dickson[58], Verónica S. Di Stilio[59], Shona Ellis[7], Eva Fast[9], Nicole Feja[9], Katie J. Field[60], Dmitry A. Filatov[61], Patrick M. Finnegan[31], Sandra K. Floyd[62], Bruno Fogliani[44,63], Nicolás García[64], Gildas Gâteblé[63], Grant T. Godden[6], Falicia (Qi Yun) Goh[65], Stephan Greiner[66], Alex Harkess[1,30], James Mike Heaney[5,6], Katherine E. Helliwell[67,68], Karolina Heyduk[1,69], Julian M. Hibberd[38], Richard G. J. Hodel[5,6,34], Peter M. Hollingsworth[28], Marc T. J. Johnson[70], Ricarda Jost[31,71], Blake Joyce[40,72], Maxim V. Kapralov[73], Elena Kazamia[38], Elizabeth A. Kellogg[30,74], Marcus A. Koch[75], Matt Von Konrat[76], Kálmán Könyves[77,78], Toni M. Kutchan[30], Vivienne Lam[7], Anders Larsson[79], Andrew R. Leitch[80], Roswitha Lentz[9], Fay-Wei Li[81], Andrew J. Lowe[82], Martha Ludwig[52], Paul S. Manos[83], Evgeny Mavrodiev[5,6], Melissa K. McCormick[84], Michael McKain[85], Tracy McLellan[86], Joel R. McNeal[1,87], Richard E. Miller[88], Matthew N. Nelson[89,90,91], Yanhui Peng[40,92], Paula Ralph[55], Daniel Real[93], Chance W. Riggins[94], Markus Ruhsam[28], Rowan F. Sage[95], Ann K. Sakai[96], Moira Scascitella[7], Edward E. Schilling[97], Eva-Marie Schlösser[9], Heike Sederoff[98], Stein Servick[5], Emily B. Sessa[5], A. Jonathan Shaw[83], Shane W. Shaw[99], Erin M. Sigel[100], Cynthia Skema[101], Alison G. Smith[38], Ann Smithson[31], C. Neal Stewart Jr[40,41], John R. Stinchcombe[95,102], Peter Szövényi[103], Jennifer A. Tate[58], Helga Tiebel[9], Dorset Trapnell[1], Matthieu Villegente[44], Chun-Neng Wang[104], Stephen G. Weller[96], Michael Wenzel[57], Stina Weststrand[105], James H. Westwood[106], Dennis F. Whigham[84], Shuangxiu Wu[107,134], Adrien S. Wulff[44,63], Yu Yang[108], Dan Zhu[109], Cuili Zhuang[7], Jennifer Zuidof[110], Mark W. Chase[26,111,135], J. Chris Pires[112,134], Carl J. Rothfels[83,113,114,134], Jun Yu[107,134], Cui Chen[115], Li Chen[116], Shifeng Cheng[117], Juanjuan Li[116], Ran Li[116], Xia Li[116], Haorong Lu[116], Yanxiang Ou[116], Xiao Sun[118], Xuemei Tan[116], Jingbo Tang[119], Zhijian Tian[115], Feng Wang[120], Jun Wang[121], Xiaofeng Wei[116], Xun Xu[116], Zhixiang Yan[116], Fan Yang[116], Xiaoni Zhong[118], Feiyu Zhou[116], Ying Zhu[116], Yong Zhang[116,118,135], Saravanaraj Ayyampalayam[1,122], Todd J. Barkman[123], Nam-phuong Nguyen[124], Naim Matasci[125], David R. Nelson[126], Erfan Sayyari[10], Eric K. Wafula[55], Ramona L. Walls[72], Tandy Warnow[127,134], Hong An[128], Nils Arrigo[2], Anthony E. Baniaga[2], Sally Galuska[2], Stacy A. Jorgensen[129], Thomas I. Kidder[2], Hanghui Kong[130], Patricia Lu-Irving[2], Hannah E. Marx[2,34], Xinshuai Qi[2], Chris R. Reardon[2], Brittany L. Sutherland[2], George P. Tiley[83], Shana R. Welles[2], Rongpei Yu[131], Shing Zhan[113], Lydia Gramzow[132], Günter Theißen[132] & Gane Ka-Shu Wong[3,116,133,135]*

[1]Department of Plant Biology, University of Georgia, Athens, GA, USA. [2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [3]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada. [4]Department of Biology, The University of British Columbia Okanagan, Kelowna, British Columbia, Canada. [5]Department of Biology, University of Florida, Gainesville, FL, USA. [6]Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. [7]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada. [8]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Germany. [9]Botanical Institute, University of Cologne, Cologne, Germany. [10]Department of Electrical and Computer Engineering, University of California, San Diego, San Diego, CA, USA. [11]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany. [12]Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany. [13]BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany. [14]Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany. [15]UF Biodiversity Institute, and UF Genetics Institute, University of Florida, Gainesville, FL, USA. [16]New York Botanical Garden, New York, NY, USA. [17]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön, Germany. [18]Negaunee Institute for Plant Conservation Science and Action, Chicago Botanic Garden, Glencoe, IL, USA. [19]Program in Plant Biology and Conservation, Northwestern University, Evanston, IL, USA. [20]Department of Horticulture, Michigan State University, East Lansing, MI, USA. [21]Department of Botany, University of Wisconsin-Madison, Madison, WI, USA. [22]Ocean University of China, Qingdao, China. [23]Department of Biological Sciences, University of North Texas, Denton, TX, USA. [24]Centre for Plant Biotechnology and Genomics (CBGP, UPM-INIA), Madrid, Spain. [25]Department of Biodiversity and Conservation, Real Jardín Botánico (RJB-CSIC), Madrid, Spain. [26]Jodrell Laboratory, Royal Botanic Gardens, Kew, London, UK. [27]School of Marine Sciences, University of Maine, Orono, ME, USA. [28]Royal Botanic Garden Edinburgh, Edinburgh, UK. [29]Department of Plant Biology, Laval University, Quebec, Quebec, Canada. [30]Donald Danforth Plant Science Center, St Louis, MO, USA. [31]School of Biological Sciences, The University of Western Australia, Perth, Western Australia, Australia. [32]Kings Park and Botanic Garden, Department of Biodiversity, Conservation and Attractions, Perth, Western Australia, Australia. [33]Australian Tropical Herbarium, James Cook University, Cairns, Queensland, Australia. [34]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA. [35]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK. [36]Umeå Plant Science Centre, Umeå Universitet, Umeå, Sweden. [37]Australian Centre for Evolutionary Biology and Biodiversity, Environment Institute, School of Earth and Environmental Science, University of Adelaide, Adelaide, South Australia, Australia. [38]Department of Plant Sciences, University of Cambridge, Cambridge, UK. [39]Royal Botanic Garden Sydney, Sydney, New South Wales, Australia. [40]Department of Plant Sciences, University of Tennessee, Knoxville, TN, USA. [41]Center for Agricultural Synthetic Biology, University of Tennessee, Knoxville, TN, USA. [42]Department of Food Science, University of Tennessee, Knoxville, TN, USA. [43]Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, NC, USA. [44]Institute for Exact and Applied Sciences, University of New Caledonia, Noumea, New Caledonia. [45]Department of Biology, University of Massachusetts, Amherst, MA, USA. [46]USDA-Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA, USA. [47]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA. [48]Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, UK. [49]Department of Plant Biology, University of Illinois, Urbana-Champaign, Urbana, IL, USA. [50]Fairy Lake Botanical Garden, Chinese Academy of Sciences, Shenzhen, China. [51]Yale-NUS College, Singapore, Republic of Singapore. [52]School of Molecular Sciences, The University of Western Australia, Perth, Western Australia, Australia. [53]Department of Plant Biology, Southern Illinois University, Carbondale, IL, USA. [54]Department of Anatomy, University of Otago, Dunedin, New Zealand. [55]Biology Department, Pennsylvania State University, University Park, PA, USA. [56]Department of Biological Science, California State University Fullerton, Fullerton, CA, USA. [57]Atlanta Botanical Garden, Atlanta, GA, USA. [58]Massey University, School of Fundamental Sciences, Palmerston North, New Zealand. [59]Department of Biology, University of Washington, Seattle, WA, USA. [60]Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds, UK. [61]Department of Plant Sciences, University of Oxford, Oxford, UK. [62]School of Biological Sciences, Monash University, Melbourne, Victoria, Australia. [63]Institut Agronomique néo-Calédonien (IAC), Equipe ARBOREAL, Païta, New Caledonia. [64]Facultad de Ciencias Forestales y de la Conservación de la Naturaleza, Universidad de Chile, Santiago, Chile. [65]Genome Institute of Singapore, Singapore, Singapore. [66]Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany. [67]Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, UK. [68]Marine Biological Association, The Laboratory, Plymouth, UK. [69]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. [70]Department of Biology, University of Toronto Mississauga, Mississauga, Ontario, Canada. [71]School of Life Sciences, La Trobe University, Bundoora, Victoria, Australia. [72]CyVerse, BIO5 Institute, University of Arizona, Tucson, AZ, USA. [73]School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, UK. [74]University of Missouri, St Louis, St Louis, MO, USA. [75]Centre for Organismal Studies Heidelberg, Department of Biodiversity and Plant Systematics, Botanic Garden and Herbarium Heidelberg, University of Heidelberg, Heidelberg, Germany. [76]The Field Museum, Chicago, IL, USA. [77]Royal Horticultural Society Garden Wisley, Woking, UK. [78]University of Reading Herbarium, School of Biological Sciences, University of Reading, Reading, UK. [79]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden. [80]School of Biological and Chemical Sciences, Queen Mary University of London, London, UK. [81]Boyce Thompson Institute, Cornell University, Ithaca, NY, USA. [82]Environment Institute, School of Biological Science, University of Adelaide, Adelaide, South Australia, Australia. [83]Department of Biology, Duke University, Durham, NC, USA. [84]Smithsonian Environmental Research Center, Edgewater, MD, USA. [85]Department of Biological Sciences, University of Alabama, Tuscaloosa, AL, USA. [86]School of Molecular and Cell Biology, University of the Witwatersrand, Johannesburg, South Africa. [87]Department of Ecology, Evolution and Organismal Biology, Kennesaw State University, Kennesaw, GA, USA. [88]Flower Diversity Institute, Arvada, CO, USA. [89]CSIRO Agriculture and Food, Perth, Western Australia, Australia. [90]Millennium Seed Bank, Wakehurst, Royal Botanic Gardens, Kew, Ardingly, UK. [91]The UWA Institute of Agriculture, The University of Western Australia, Perth, Western Australia, Australia. [92]Centers for Disease Control and Prevention, Atlanta, GA, USA. [93]Department of Primary Industries and Regional Development, Perth, Western Australia, Australia. [94]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [95]Department of Ecology and Evolutionary Biology, The University of Toronto, Ontario, Canada. [96]Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA, USA. [97]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, USA. [98]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, USA. [99]Manoa, Honolulu, HI, USA. [100]Department of Biology, University of Louisiana at Lafayette, Lafayette, LA, USA. [101]Morris Arboretum of the University of Pennsylvania, Philadelphia, PA, USA. [102]Koffler Scientific Reserve, University of Toronto, King City, Ontario, Canada. [103]Department of Systematic and Evolutionary Botany, University of Zurich, Zurich, Switzerland. [104]National Taiwan University, Institute of Ecology and Evolutionary Biology, Department of Life Science, Taipei, Taiwan. [105]Systematic Biology, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden. [106]Department of Plant Pathology, Physiology and Weed Science, Virginia Tech, Blacksburg, VA, USA. [107]CAS Key Laboratory of Genome Sciences and Information, Beijing Key Laboratory of Genome and Precision Medicine Technologies, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. [108]Key Laboratory of Agricultural Biological Functional Genes, Northeast Agricultural University, Harbin, China. [109]College of Life Science, Qingdao Agricultural University, Qingdao, China. [110]Agriculture and Agri-Food Canada, Lacombe, Alberta, Canada. [111]Department of Environment and Agriculture, Curtin University, Bentley, Western Australia, Australia. [112]Bond Life Sciences Center, Division of Biological Sciences, University of Missouri, Columbia, MO, USA. [113]Department of Zoology, University of British Columbia, Vancouver, British

Columbia, Canada. [114]University Herbarium and Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA. [115]Beijing Genomics Institute-Wuhan, Wuhan, China. [116]BGI-Shenzhen, Shenzhen, China. [117]Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [118]Huahan Gene, Shenzhen, China. [119]MGI, BGI-Shenzhen, Shenzhen, China. [120]Allwegene Technology, Beijing, China. [121]iCarbonX, Shenzhen, China. [122]Georgia Advanced Computing Resource Center, University of Georgia, Athens, GA, USA. [123]Department of Biological Sciences, Western Michigan University, Kalamazoo, MI, USA. [124]Department of Computer Science and Engineering, University of California, San Diego, San Diego, CA, USA. [125]Lawrence J. Ellison Institute for Transformative Medicine, University of Southern California, Los Angeles, CA, USA. [126]Microbiology, Immunology and Biochemistry, The University of Tennessee Health Science Center, Memphis, TN, USA. [127]Department of Computer Science, University of Illinois, Urbana-Champaign, Urbana, IL, USA. [128]Division of Biological Sciences, University of Missouri, Columbia, MO, USA. [129]Arizona Research Laboratories, University of Arizona, Tucson, AZ, USA. [130]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China. [131]Flower Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, China. [132]Department of Genetics, Matthias Schleiden Institute, Friedrich-Schiller-University Jena, Jena, Germany. [133]Department of Medicine, University of Alberta, Edmonton, Alberta, Canada. [134]These authors contributed equally: Eric J. Carpenter, Matthew A. Gitzendanner, Zheng Li, Siavash Mirarab, Martin Porsch, Kristian K. Ullrich, Lisa DeGironimo, Patrick P. Edger, Ingrid E. Jordon-Thaden, Steve Joya, Tao Liu, Barbara Melkonian, Nicholas W. Miles, Lisa Pokorny Montero, Charlotte Quigley, Philip Thomas, Juan Carlos Villarreal. [135]These authors jointly supervised this work: James H. Leebens-Mack, Michael S. Barker, Michael K. Deyholos, Sean W. Graham, Ivo Grosse, Michael Melkonian, Siavash Mirarab, Marcel Quint, Stefan A. Rensing, Douglas E. Soltis, Pamela S. Soltis, Dennis W. Stevenson, Claude W. dePamphilis, Mark W. Chase, J. Chris Pires, Carl J. Rothfels, Jun Yu, Yong Zhang, Tandy Warnow, Shuangxiu Wu, Gane Ka-Shu Wong. *e-mail: jleebensmack@uga.edu; gane@ualberta.ca

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, although simulations included in the genome duplication analyses did include drawing from random distributions. The investigators were not blinded to allocation during experiments and outcome assessment.

Transcriptome sequencing

RNA was isolated from young vegetative tissue from all of the species that were included in our phylogenomic analyses as described elsewhere[39,67,68]. Reproductive tissues were also included for some species (Supplementary Table 1). Transcript assembly, contaminant identification and gene-family circumscription were also performed as described previously[39] and are described in more detail in the Supplementary Methods.

Phylogeny reconstruction

Analyses were performed on single-copy gene trees using ASTRAL to account for variation among gene trees owing to incomplete lineage sorting[15,69]. ASTRAL analyses were performed on gene trees estimated from unbinned amino acid alignments, first and second codons, statistically binned supergenes with unweighted bins[70,71] and filtered taxon sets (excluding 'rogue' taxa as described below), with filtering of gene-tree bootstrap support thresholds of up to 33% to see whether the effects of gene-tree estimation error could be reduced (Supplementary Fig. 6). Binning left the majority of genes in singleton bins and had minimal effects on the overall species tree. Unless otherwise specified, we use 'ASTRAL topology' to refer to the tree inferred from 410 unbinned amino acid alignments in which branches with 33% or less support are contracted. In addition, supermatrix analyses were performed on concatenated nuclear gene alignments and concatenated plastid gene alignments compiled using previously described methods[72]. All scripts used to perform analyses on the nuclear gene data are available at https://doi.org/10.5281/zenodo.3255100.

Multiple sequence alignment and data filtering. We built a multiple sequence alignment based on predicted amino acid sequences of each gene and forced DNA sequences to conform to the amino acid alignment. We first divided sequences in each gene into two subsets, full-length and abnormal sequences, and then used PASTA[73] with default settings to align full-length sequences and UPP[74] to add abnormal sequences to the full-length alignment. We designated as abnormal any sequence that was 66% shorter or 66% longer than the median length of the full-length gene sequences. Once UPP alignments were obtained, we removed from them all unaligned (that is, insertion) sites. DNA alignments were then derived from amino acid sequence alignments (FAA2FNA) and third codon positions were removed owing to extreme among-species variation in GC content (Supplementary Fig. 7). To reduce running time, we then masked all sites from the alignment that contained more than 90% gaps. Finally, because the inclusion of fragmentary data in gene-tree estimation can be problematic[75], we removed any sequence that had a gap for at least 67% of the sites in the site-filtered alignment (the 67% threshold was chosen based on simulation results[75]). Gene sequence occupancy for 410 single-copy genes in the 1,178 accessions used in our analyses is displayed as a frequency histogram (Supplementary Fig. 4) and a heat map (Supplementary Fig. 5).

In addition to filtering gappy sites and fragmentary sequences, we identified and removed sequences that were placed on extremely long branches on their respective gene trees. To identify these, we used the initial alignments to build gene trees (see below). We then rooted each gene tree by finding the bipartition that separated the largest exclusive group of outgroup or red algae taxa. If red algae were entirely missing for the gene, we used Glaucophyta, Prasinococcales, prasinophytes, *Volvox carteri*, *Chlamydomonas reinhardtii* or *Klebsormidium nitens*. We then removed any sequences that had a root-to-tip distance that was four standard deviations longer than the median root-to-tip distance in each gene tree. Once these sequences on long branches were removed, alignments were re-estimated using the same approach described above, and new gene trees were estimated.

Gene-tree estimation. To estimate gene trees, we used RAxML v.8.1.17[76], with one starting tree for building initial trees (used for long-branch filtering) and 10 different starting trees for final gene trees. Support was assessed with 100 replicates of bootstrapping. For DNA analyses, the GTR substitution model and the GAMMA-distributed site rates were used. For amino acid sequences, we used a Perl script adapted from the RAxML website to search among 16 different substitution models on a fixed starting tree per gene and chose the model with the highest likelihood (JTT, JTTF or JTTDCMUT were selected for 349 out of 410 genes). For amino acid trees, we also used the GAMMA-distributed site rates.

Species tree estimation. We used ASTRAL-II[15] v.5.0.3 to estimate the species tree on the basis of all 410 genes; using 384 genes that each included at least half of the species changed only 3 low-support branches. We used multi-locus bootstrapping[77,78] and the built-in local posterior probabilities of ASTRAL to estimate branch support[69] and to test for polytomies[79], drawn on species trees estimated based on the maximum-likelihood gene trees. We also used the built-in functionality of ASTRAL (version 4.11.2) to compute the percentage of gene trees that agreed with each branch in the species tree, by finding the average number of gene-tree quartets defined around the branch (choosing one taxon from each side) that were congruent with the species tree and used DiscoVista[80] to visualize them (Fig. 4). Median representation of each species across the 410 single-copy gene trees was 82.4% with 88.2% and 67.1% of species having assemblies for at least 50% or 75% of the 410 single-copy genes, respectively. A large body of work on phylogenetic methodologies has established that gene and species tree estimation can be robust to missing data, particularly with dense taxon sampling[75,81,82]. Recent papers have even established statistical consistency under missing data[83]. Similar evidence of robustness also exists in the context of concatenated analyses[84–86].

All supermatrix analyses are based on the filtered amino acid and first and second codon position alignments that included at least half of the species for 384 genes. The (1) unfiltered supermatrices used the gene alignments as is; the (2) eudicot supermatrices retained only eudicot species in the supermatrix; and the (3) supermatrices with eight 'rogue' taxa removed (*Dillenia indica*, *Tetrastigma obtectum*, *Tetrastigma voinierianum*, *Vitis vinifera*, *Cissus quadrangularis*, '*Spirotaenia*' sp., *Ceratophyllum demersum* and *Prasinococcus capsulatus*) that varied in placement among our full ASTRAL, supermatrix and plastid genome analyses. Well-supported branching orders were stable among analyses (Supplementary Fig. 6).

Maximum-likelihood supermatrix analyses were performed using ExaML v.3.0.14[87]. Similar to the gene-tree analyses, the GAMMA model of rate heterogeneity across sites was used for all maximum-likelihood supermatrix analyses. To better handle model heterogeneity across genes, we divided the supermatrix into partitions. For the amino acid alignments, the protein model selected for each gene family in the gene-tree estimation process was used to group genes into partitions, creating one partition per substitution model. For the nucleotide alignments, we estimated the GTR transition rate parameters and the alpha shape parameter for each codon position (first and second positions) of each alignment using RAxML v.8.1.21[76]. We then projected the maximum-likelihood parameter values for each gene into a two-dimensional plane using principal component analysis[88]. We performed *k*-means clustering[89] in R[90] to group the codon positions into partitions, selecting $k = 8$, which accounted for 80% of the variation. Trees derived from nucleotide alignments can be found at https://doi.org/10.5281/zenodo.3255100).

To examine the influence of the starting tree on the likelihood of the final tree, we performed preliminary analyses on an earlier version of our supermatrices. We generated nine different maximum-parsimony starting trees using RAxML v.8.1.21 and one maximum-likelihood starting tree using FastTree-2 v.2.1.5[91]. We then ran ExaML on each of the starting trees, noting the final maximum-likelihood score. We found that in all cases, the ExaML maximum-likelihood tree using the FastTree-2 maximum-likelihood starting tree had a better maximum-likelihood score than any of the ExaML maximum-likelihood trees using maximum-parsimony starting trees. Thus, for all of the supermatrix analyses, we used FastTree-2 to generate our initial starting tree. Support was inferred for the branches of the final tree from 100 bootstrap replicates.

Outgroup taxa from outside Archaeplastida were used to root all species trees estimated using nuclear genes (all ASTRAL and supermatrix analyses). The plastome supermatrix tree for Viridiplantae was rooted using Rhodophyta as outgroup.

Inferring and placing WGDs

DupPipe analyses of WGDs from transcriptomes of single species. For each transcriptome, we used the DupPipe pipeline to construct gene families and estimate the age distribution of gene duplications[16,17]. We translated DNA sequences and identified reading frames by comparing the Genewise[92] alignment to the best-hit protein from a collection of proteins from 25 plant genomes from Phytozome[93]. For all DupPipe runs, we used protein-guided DNA alignments to align our nucleic acid sequences while maintaining the reading frame. We estimated synonymous divergence (K_s) using PAML with the F3X4 model[94] for each node in the gene-family phylogenies. We identified peaks of gene duplication as evidence of ancient WGDs in histograms of the age distribution of gene duplications (K_s plots). We identified species with potential WGDs by comparing their paralogue age distribution to a simulated null using a Kolmogorov–Smirnov goodness of fit test[95]. We then used mixture modelling and manual curation to identify significant peaks consistent with a potential WGD and to estimate their median paralogue K_s values. Significant peaks were identified using a likelihood ratio test in the boot.comp function of the package mixtools in R[96].

Estimating orthologous divergence. To place putative WGDs in relation to lineage divergence, we estimated the synonymous divergence of orthologues among pairs of species that may share a WGD based on their phylogenetic position and evidence from the within-species K_s plots. We used the RBH Orthologue pipeline[17] to estimate the mean and median synonymous divergence of orthologues and compared those to the synonymous divergence of inferred paleopolyploid peaks. We identified orthologues as reciprocal best blast hits in pairs of transcriptomes. Using protein-guided DNA alignments, we estimated the pairwise synonymous divergence for each pair of orthologues using PAML with the F3X4 model[94]. WGDs were interpreted to have occurred after lineage divergence if the median synonymous divergence of WGD paralogues was younger than the median synonymous divergence of orthologues. Similarly, if the synonymous divergence of WGD paralogues was older than that orthologue synonymous divergence, then we interpreted those WGDs as shared.

MAPS analyses of WGDs from transcriptomes of multiple species. To infer and locate putative WGDs in our datasets, we used a gene-tree sorting and counting algorithm, the multi-taxon paleopolyploidy search (MAPS) tool[18]. For each MAPS analysis, we selected at least two species that potentially share a WGD in their ancestry as well as representative species from lineages that may phylogenetically bracket the WGD. MAPS uses this given species tree to filter collections of nuclear gene trees for subtrees consistent with relationships at each node in the species tree. Using this filtered set of subtrees, MAPS identifies and records nodes with a gene duplication shared by descendant taxa. To infer and locate a potential WGD, we compared the number of duplications observed at each node to a null simulation of background gene birth and death rates[97,98]. A Fisher's exact test, implemented in R[90], was used to identify locations with significant increases in gene duplication compared with a null simulation (Supplementary Table 5). Locations with significantly more duplications than expected were then compared to a simulated WGD at this location. If the observed duplications were consistent with this simulated WGD using Fisher's exact test, we identified the location as a WGD if it was consistent with inferences from K_s plots and orthologue divergence data. In some cases, MAPS inferred significant duplications without apparent signatures in K_s plots or previously published research. In these cases, we recognized the event as a significant burst of gene duplication.

Each MAPS analysis was designed to place focal WGDs near the centre of a species tree to minimize errors in WGD inference. Errors in transcriptome or genome assembly, gene-family clustering and the construction of gene-family phylogenies can result in topological errors in gene trees[99]. Previous studies have suggested that errors in gene trees can lead to biased placements of duplicates towards the root of the tree and losses towards the tips of the tree[100]. For this reason, we aimed to put focal nodes for a particular MAPS analysis test in the middle of the phylogeny. To further decrease potential error in our inferences of gene duplications, we required at least 45% of the ingroup taxa to be present in all subtrees analysed by MAPS[97]. If this minimum requirement of ingroup taxa numbers is not met, the gene subtree will be filtered out and excluded from our analysis. Increasing taxon occupancy leads to a more accurate inference of duplications and reduces some of the biases in mapping duplications onto a species tree[100,101]. To maintain sufficient gene-tree numbers for each MAPS analysis, we used collections of gene-family phylogenies for six to eight taxa to infer ancient WGDs.

For each MAPS analysis, the transcriptomes were translated into amino acid sequences using the TransPipe pipeline[17]. Using these translations, we performed reciprocal protein BLAST (BLASTp) searches among datasets for the MAPS analysis using a cut-off of $E = 1 \times 10^{-5}$. We clustered gene families from these BLAST results using OrthoFinder under the default parameters[102]. Using a custom Perl script (https://bitbucket.org/barkerlab/MAPS), we filtered for gene families that contained at least one gene copy from each taxon in a given MAPS analysis and discarded the remaining OrthoFinder clusters. We used PASTA[73] for automatic alignment and phylogeny reconstruction of gene families. For each gene-family phylogeny, we ran PASTA until we reached three iterations without an improvement in likelihood score using a centroid breaking strategy. Within each iteration of PASTA, we constructed subset alignments using MAFFT[103], used Muscle[104] for merging these subset alignments and RAxML[76] for tree estimation. The parameters for each software package were the default options for PASTA (https://bitbucket.org/barkerlab/1kp). We used the best-scoring PASTA tree for each multi-species nuclear gene family to collectively estimate the numbers of shared gene duplications on each branch of the given species.

To generate null simulations, we first estimated the mean background gene duplication rate (λ) and gene loss rate (μ) with WGDgc[98] (Supplementary Tables 5, 11). Gene count data were obtained from OrthoFinder[102] clusters associated with each species tree (Supplementary Table 5). λ and μ were estimated using only gene clusters that spanned the root of their respective species trees, which has been shown to reduce biases in the maximum-likelihood estimates[98] of λ and μ. We chose a maximum gene-family size of 100 for parameter estimation, which was necessary to provide an upper bound for numerical integration of node states[98]. We provided a prior probability distribution on the number of genes at the root of each species tree, such that ancestral gene-family sizes followed a shifted geometric distribution with mean equal to the average number of genes per gene family across species (Supplementary Table 5).

Gene trees were then simulated within each MAPS species trees using the GuestTreeGen program from GenPhyloData[105]. For each species tree, we simulated 3,000 gene trees with at least one tip per species: 1,000 gene trees at the λ and μ maximum-likelihood estimates, 1,000

gene trees at half the estimated λ and μ, and 1,000 trees at three times λ and μ. For all simulations, we applied the same empirical prior used for estimation of λ and μ. We then randomly resampled 1,000 trees without replacement from the total pool of gene trees 100 times to provide a measure of uncertainty on the percentage of subtrees at each node. For positive simulations of WGDs, we simulated gene trees using the same approach used to generate null distributions (Supplementary Table 5) but incorporated a WGD at the test branch. Previous empirical estimates of paralogues retained following a plant WGD are 10% on average[106]. To be conservative for inferring WGDs in our MAPS analyses, we allowed at least 20% of the genes to be retained following the simulated WGD to account for biased gene retention and loss. For WGDs that might have a lower gene retention rate, we used an additional simulation using 15% gene retention (Supplementary Table 6).

Gene-family evolution

Transcriptome-based gene-family size estimation. To robustly estimate gene-family sizes from transcriptomic data, we needed to overcome three major challenges: (1) the fragmentation of transcript sequences; (2) the absence of low-abundance transcripts; and (3) the over-prediction of gene-family sizes due to assembly duplications and biological isoforms. We dealt with these challenges as follows.

Fragmentation of data. The multiple sequence alignments used to construct the domain-specific profile hidden Markov models (HMMs) ranged from 23 to 463 amino acids in length; 78% of these alignments were shorter than 120 amino acids, and 84.6% of the assembled and translated transcripts were longer than 120 amino acids. By mainly characterizing gene families using single domains (Supplementary Table 9), we limited the effect of the fragmentation of transcripts from the assembly of short read data. HMMs used for gene-family classification and decision rules obtained from either published work[107] or gene-family experts are given in Supplementary Table 9; 12 out of 23 gene families were classified by a single 'should' rule, 2 out of 23 were defined by a XOR 'should' rule, which also classifies a sequence by the presence of a single domain, 8 out of 23 gene families were classified by a more complex rule set including 'should not' rules. The only gene family for which multiple domains needed to be present was the PLS subfamily of the PPR gene family.

Loss of low abundance transcripts. To account for possible bias in the sampling of the gene space, all species that showed low levels of transcriptome completeness were removed. The lowest value of transcriptome completeness obtained from 30 annotated plant genomes was used as the lower exclusion limit. We removed all samples in which more than 42.5% of BUSCO[31] sequences were missing using default settings and the eukaryotic dataset as the query database.

Gene-family over-prediction. We clustered assembled protein sequences by sequence similarity and merged sequences that showed at least 99% identity. To check for the possibility of merging sequences that should be counted separately, different identity cut-offs were compared between the 1KP datasets and 32 annotated plant genomes.

Extended Data Figure 3c, d shows the average gene-family sizes for 23 gene families and 13 clades obtained from 1KP samples and 32 annotated plant genomes. These gene-family sizes show a high Pearson correlation ($r = 0.95$) between 1KP samples and plant genomes, and therefore a linear relationship between the two approaches is indicated. The results from the 1KP dataset are on average smaller by a factor of 2.3. Although this is a clear underestimate, the scale factor by which the estimate is too small is relatively consistent, especially as the gene-family sizes increase.

Sequence clustering. We used cdhit v.4.5.7[108,109] to reduce the number of protein sequence duplications in the dataset. We assessed 100%, 99.5%, 99%, 95% and 90% sequence identity thresholds. The percentage of remaining sequences for the 1KP samples and 32 reference genomes is displayed in Extended Data Fig. 3f. We chose 99% sequence identity as the value to use for this study.

Estimation of gene-family size. Gene-family experts provided the knowledge to classify protein sequences as members of gene families with profile HMMs. In total, 46 HMMs representing 23 large gene families[30] were used to estimate gene-family sizes in the analysed species. Classification rules and HMMs for 14 gene families that have been published previously[107] were converted to HMMER3 format and used in this study. Gene-family classification rules and HMMs for the remaining nine families can be found in Supplementary Table 8. HMMs were taken from the Pfam database (accessed 12 May 2016) or were provided by gene-family experts (Supplementary Table 8). HMMER[110] (v.3.1b2) was used to scan for matches in the filtered 1KP dataset. Where available, gathering thresholds were used; otherwise an *E*-value cut-off was applied to indicate domain presence. If the *E* value is not noted in Supplementary Table 9, the default *E* value of 10 was applied. The results on the species level are listed in Supplementary Table 10s.

Statistical test for expansions and contractions. To assess whether a gene family expanded or contracted in a lineage, we compared a weighted average of gene numbers in adjacent clades and grades (Fig. 4). We also checked for expansions and contractions within clades but did not find any statistically significant shifts. The counts of gene-family members from two clades or grades were compared with a Kolmogorov–Smirnov test with a *P*-value threshold of 1×10^{-6} in R[90]. The tests conducted in this study are listed in Supplementary Table 7. Fold changes were computed using the trimmed arithmetic mean in which the top and bottom 5% of the data were discarded. Only expansions larger than 1.5 fold (or contractions smaller than 2/3) are reported.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All raw sequence reads have been posted in the NCBI SRA database under BioProject accession PRJEB4922. SRA entries for each assembly are listed in Supplementary Table 1. All sequence, gene tree and species tree data can be accessed through CyVerse Data Commons at https://doi.org/10.25739/8m7t-4e85. In addition, gene-family nucleotide and amino acid FASTA files can also be found at http://jlmwiki.plantbio.uga.edu/onekp/v2/; multiple sequence alignments, gene trees and species trees for single-copy nuclear genes included in phylogenomic analyses are also at https://doi.org/10.5281/zenodo.3255100; K_s plots, alignments and trees used for WGD analyses can be found at https://bitbucket.org/barkerlab/1kp; and data used for gene-family expansion analyses can be found at https://github.com/GrosseLab/OneKP-gene-family-evo.

Code availability

Scripts used for phylogenomic species tree analyses are available at https://doi.org/10.5281/zenodo.3255100. Scripts used for MAPS analyses of WGDs are available at https://bitbucket.org/barkerlab/maps and scripts used for gene-family expansion analyses are available at https://github.com/GrosseLab/OneKP-gene-family-evo. All script files are also accessible through CyVerse Data Commons at https://doi.org/10.25739/8m7t-4e85.

67. Johnson, M. T. J. et al. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLoS ONE* **7**, e50226 (2012).
68. Jordon-Thaden, I. E., Chanderbali, A. S., Gitzendanner, M. A. & Soltis, D. E. Modified CTAB and TRIzol protocols improve RNA extraction from chemically complex Embryophyta. *Appl. Plant Sci.* **3**, 1400105 (2015).

69. Sayyari, E. & Mirarab, S. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).
70. Mirarab, S., Bayzid, M. S., Boussau, B. & Warnow, T. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* **346**, 1250463 (2014).
71. Bayzid, M. S., Mirarab, S., Boussau, B. & Warnow, T. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLoS ONE* **10**, e0129183 (2015).
72. Gitzendanner, M. A., Soltis, P. S., Wong, G. K.-S., Ruhfel, B. R. & Soltis, D. E. Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *Am. J. Bot.* **105**, 291–301 (2018).
73. Mirarab, S., Nguyen, N. & Warnow, T. PASTA: ultra-large multiple sequence alignment. *Res. Comput. Mol. Biol.* **8394**, 177–191 (2014).
74. Nguyen, N.-P. D., Mirarab, S., Kumar, K. & Warnow, T. Ultra-large alignments using phylogeny-aware profiles. *Genome Biol.* **16**, 124 (2015).
75. Sayyari, E., Whitfield, J. B. & Mirarab, S. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. *Mol. Biol. Evol.* **34**, 3279–3291 (2017).
76. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
77. Seo, T.-K., Kishino, H. & Thorne, J. L. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proc. Natl Acad. Sci. USA* **102**, 4436–4441 (2005).
78. Mirarab, S., Bayzid, M. S. & Warnow, T. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst. Biol.* **65**, 366–380 (2016).
79. Sayyari, E. & Mirarab, S. Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes* **9**, 132 (2018).
80. Sayyari, E., Whitfield, J. B. & Mirarab, S. DiscoVista: interpretable visualizations of gene tree discordance. *Mol. Phylogenet. Evol.* **122**, 110–115 (2018).
81. Dobrin, B. H., Zwickl, D. J. & Sanderson, M. J. The prevalence of terraced treescapes in analyses of phylogenetic data sets. *BMC Evol. Biol.* **18**, 46 (2018).
82. Molloy, E. K. & Warnow, T. To include or not to include: the impact of gene filtering on species tree estimation methods. *Syst. Biol.* **67**, 285–303 (2018).
83. Nute, M., Chou, J., Molloy, E. K. & Warnow, T. The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics* **19**, 286 (2018).
84. Wiens, J. J. Missing data, incomplete taxa, and phylogenetic accuracy. *Syst. Biol.* **52**, 528–538 (2003).
85. Wiens, J. J. & Morrill, M. C. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst. Biol.* **60**, 719–731 (2011).
86. Lam, V. K. Y. et al. Phylogenomic inference in extremis: a case study with mycoheterotroph plastomes. *Am. J. Bot.* **105**, 480–494 (2018).
87. Kozlov, A. M., Aberer, A. J. & Stamatakis, A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31**, 2577–2579 (2015).
88. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Phil. Mag.* **2**, 559–572 (1901).
89. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: a *K*-means clustering algorithm. *J. R. Stat. Soc. C* **28**, 100–108 (1979).
90. R Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org/ (R Foundation for Statistical Computing, 2014).
91. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
92. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
93. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
94. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
95. Cui, L. et al. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).
96. Benaglia, T., Chauveau, D., Hunter, D. & Young, D. mixtools: an r package for analyzing mixture models. *J. Stat. Softw.* **32**, 1–29 (2009).
97. Li, Z. et al. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl Acad. Sci. USA* **115**, 4713–4718 (2018).
98. Rabier, C.-E., Ta, T. & Ané, C. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Mol. Biol. Evol.* **31**, 750–762 (2014).
99. Yang, Y. & Smith, S. A. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* **14**, 328 (2013).
100. Hahn, M. W. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* **8**, R141 (2007).
101. Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 150 (2015).
102. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
103. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
104. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
105. Sjöstrand, J., Arvestad, L., Lagergren, J. & Sennblad, B. GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics* **14**, 209 (2013).
106. Tiley, G. P., Ané, C. & Burleigh, J. G. Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome Biol. Evol.* **8**, 1023–1037 (2016).
107. Lang, D. et al. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* **2**, 488–503 (2010).
108. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
109. Li, W., Jaroszewski, L. & Godzik, A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **18**, 77–82 (2002).
110. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
111. Henschel, K. et al. Two ancient classes of MIKC-type MADS-box genes are present in the moss *Physcomitrella patens*. *Mol. Biol. Evol.* **19**, 801–814 (2002).

Acknowledgements The 1KP initiative was funded by the Alberta Ministry of Advanced Education and Alberta Innovates AITF/iCORE Strategic Chair (RES0010334) to G.K.-S.W., Musea Ventures, The National Key Research and Development Program of China (2016YFE0122000), The Ministry of Science and Technology of the People's Republic of China (2015BAD04B01/2015BAD04B03), the State Key Laboratory of Agricultural Genomics (2011DQ782025) and the Guangdong Provincial Key Laboratory of core collection of crop genetic resources research and application (2011A091000047). Sequencing activities at BGI were also supported by the Shenzhen Municipal Government of China (CXZZ20140421112021913/JCYJ20150529150409546/JCYJ20150529150505656). Computation support was provided by the China National GeneBank (CNGB), the Texas Advanced Computing Center (TACC), WestGrid and Compute Canada; considerable support, including personnel, computational resources and data hosting, was also provided by the iPlant Collaborative (CyVerse) funded by the National Science Foundation (DBI-1265383), National Science Foundation grants IOS 0922742 (to C.W.d., P.S.S., D.E.S. and J.H.L.-M.), IOS-1339156 (to M.S.B.), DEB 0830009 (to J.H.L.-M., C.W.d., S.W.G. and D.W.S.), EF-0629817 (to S.W.G. and D.W.S.), EF-1550838 (to M.S.B.), DEB 0733029 (to T.W. and J.H.L.-M.), and DBI 1062335 and 1461364 (to T.W.), a National Institutes of Health Grant 1R01DA025197 (to T.M.K., C.W.d. and J.H.L.-M.), Deutsche Forschungsgemeinschaft grants Qu 141/5-1, Qu 141/6-1, GR 3526/7-1, GR 3526/8-1 (to M.Q. and I.G.) and a Natural Sciences and Engineering Research Council of Canada Discovery grant (to S.W.G.). We thank all national, state, provincial and regional resource management authorities, including those of province Nord and province Sud of New Caledonia, for permitting collections of material for this research.

Author contributions Framing of research and writing was carried out by J.H.L.-M., M.S.B., E.J.C., M.K.D., M.A.G., S.W.G., I.G., Z.L., M. Melkonian, S.M., M.P., M.Q., S.A.R., D.E.S., P.S.S., D.W.S., K.K.U., N.J.W. and G.K.-S.W. Samples were collected and RNA was prepared by L.D., P.P.E., I.E.J.-T., S.J., T.L., B.M., N.W.M., L.P., C.Q., P.T., J.C.V., M.M.A., M.S.B., M.D.B., R.S.B., D.J.B., R.M.B., E.B., S.F.B., D.O.B., J.N.B., K.P.B., V.B.-S., A.L.C., S.B.C., Z.Ç., Y.C., C. Chater, J.M.C., T.C., N.D.C., H. Clayton, S. Covshoff, B.J.C.-S., H. Cross, C.W.d., J.P.D., R.D., R.C.D., V.S.D.S., S.E., E.F., N.F., K.J.F., D.A.F., P.M.F., S.K.F., B.F., N.G., G.G., M.A.G., G.T.G., F.Q.Y.G., S. Greiner, A.H., J. M. Heaney, K.E.H., K.H., J. M. Hibberd, R.G.J.H., P.M.H., M.T.J.J., R.J., B.J., M.V.K., E.K., E.A.K., M.A.K., M.V.K., K.K., T.M.K., V.L., A.L., A.R.L., R. Lentz, F.-W.L., A.J.L., M.L., P.S.M., E.M., M.K.M., M. McKain, T.M., J.R.M., R.E.M., M.N.N., Y.P., P.R., D.R., C.W.R., M.R., R.F.S., A.K.S., M.S., E.E.S., E.-M.S., H.S., S.S., E.B.S., A.J.S., S.W.S., E.M.S., C.S., A.G.S., A.S., C.N.S., J.R.S., P.S., J.A.T., H.T., D.T., M.V., C.-N.W., S.G.W., M.W., S. Weststrand, J.H.W., D.F.W., N.J.W., S. Wu, A.S.W., Y.Y., D.Z., C.Z., J.Z., M.W.C., M.K.D., S.W.G., J.H.L.-M., M. Melkonian, J.C.P., C.J.R., D.E.S., P.S.S., D.W.S. and J.Y. (jointly led by M.W.C., M.K.D., S.W.G., J.H.L.-M., M. Melkonian, J.C.P., C.J.R., D.E.S., P.S.S., D.W.S. and J.Y.; major contributions by L.D., P.P.E., I.E.J.-T., S.J., T.L., B.M., N.W.M., L.P., C.Q., P.T. and J.C.V.). RNA sequencing and transcriptome assembly were carried out by E.J.C., C. Chen, L.C., S. Cheng, J.L., R. Li, X.L., H.L., Y.O., X.S., X.T., J.T., Z.T., F.W., J.W., X.W., G.K.-S.W., X.X., Z.Y., F.Y., X.Z., F.Z., Y. Zhu and Y. Zhang. (led by Y. Zhang; major contributions by E.J.C.). Samples were validated and contaminants were filtered by J.Y., S.A., M.S.B., T.J.B., E.J.C., S.W.G., J.H.L.-M., T.L., S.M., N.-p.N., X.S., K.K.U. and S. Wu. (led by S. Wu.; major contributions by J.Y.). Gene-family circumscription and phylogenetic analyses were carried out by S.M., N.-p.N., M.A.G., S.A., J.P.D., N.M., D.R.N., E.S., D.E.S., P.S.S., D.W.S., E.K.W., R.L.W., N.J.W., C.W.d., S.W.G., J.H.L.-M. and T.W. (jointly led by S.M., C.W.d., S.W.G., J.H.L.-M. and T.W.; major contributions by: S.M.). Genome duplication analyses were carried out by Z.L., H.A., N.A., A.E.B., S. Galuska, S.A.J., T.I.K., H.K., P.L.-I., H.E.M., X.Q., C.R.R., E.B.S., B.L.S., G.P.T., S.R.W., R.Y., S.Z. and M.S.B. (led by M.S.B.; major contributions by Z.L.). Gene-family expansion analyses were carried out by M.P., K.K.U., L.G., M. Melkonian, D.R.N., G.T., G.K.-S.W., I.G., S.A.R. and M.Q. (jointly led by I.G., S.A.R. and M.Q.; major contributions by M.P. and K.K.U.).

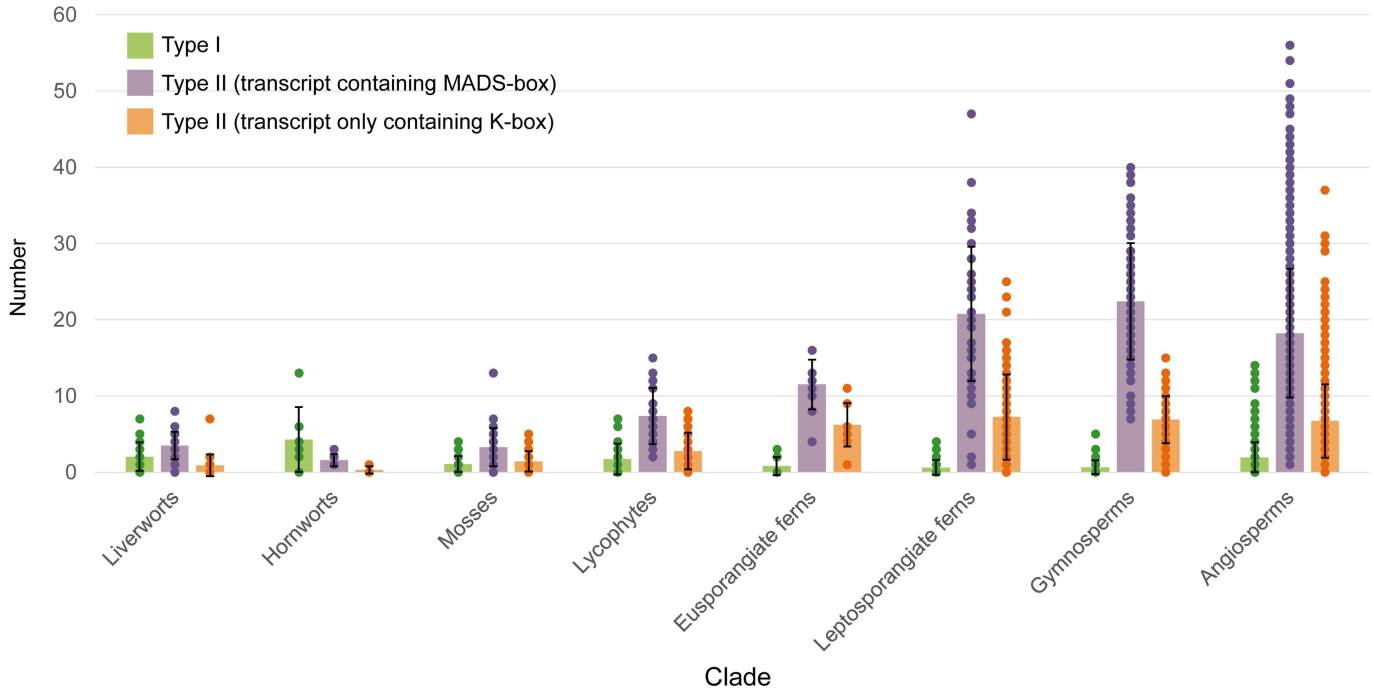
Competing interests The authors declare no competing interests.

Additional information

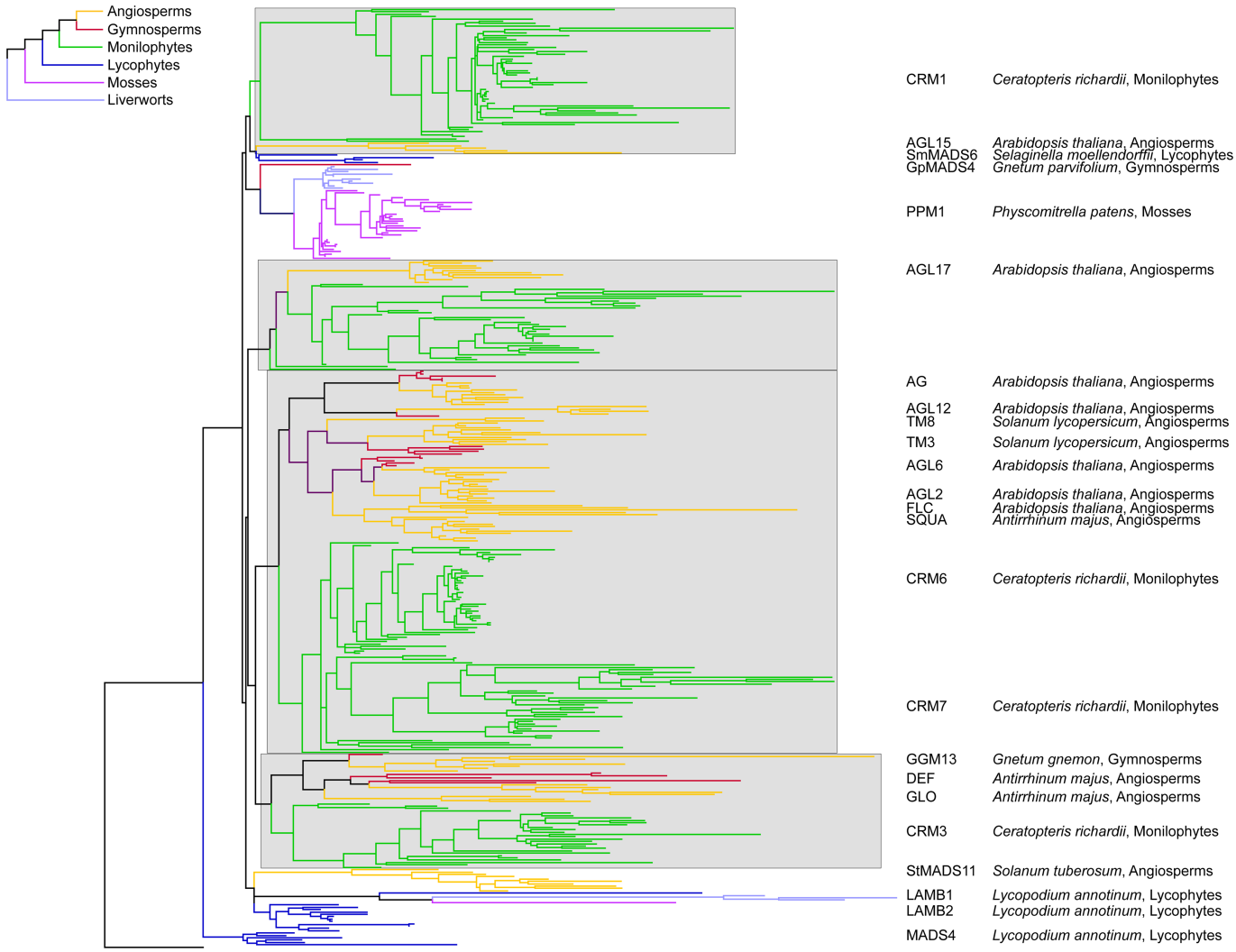
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1693-2.

Correspondence and requests for materials should be addressed to J.H.L.-M. or G.K.-S.W.

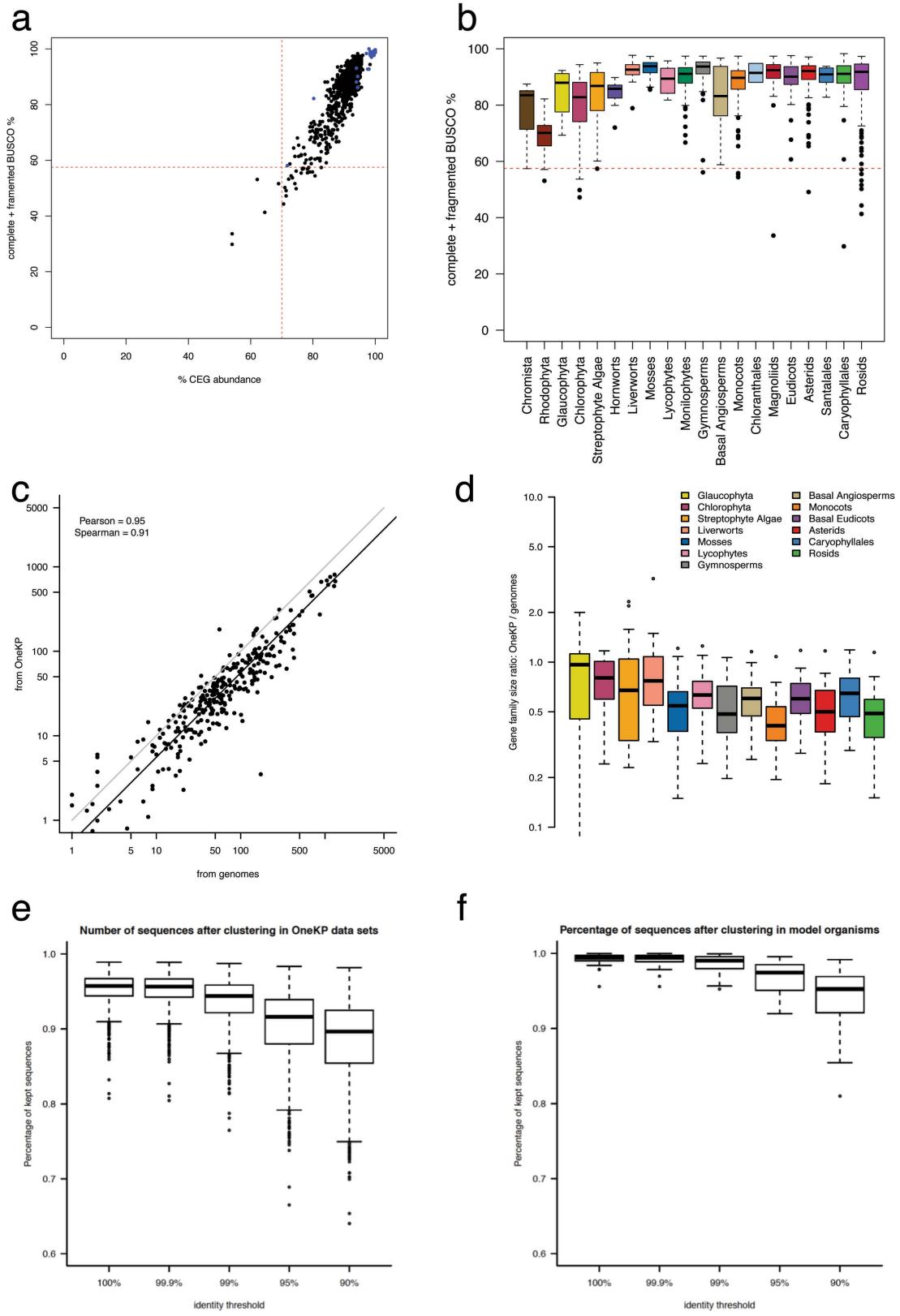
Reviewer information *Nature* thanks Paul Kenrick, Magnus Nordborg, Patrick Wincker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Mean number of MADS-box genes in the transcriptomes of different plant clades. Type I genes are shown in green; type II genes are shown in purple and orange. Transcripts in which only a K-box was identified (which are probably partial transcripts of type II genes) are shown in orange. Data are mean ± s.d. Dots indicate the numbers of MADS-box genes in individual transcriptomes. Sample sizes (n) are as follows: liverworts, $n = 26$; hornworts, $n = 7$; mosses, $n = 37$; lycophytes, $n = 22$; eusporangiate ferns, $n = 10$; leptosporangiate ferns, $n = 62$; gymnosperms, $n = 84$; and angiosperms, $n = 820$. A total of 1,068 transcriptomes were analysed for this figure.



Extended Data Fig. 2 | RAxML phylogeny of classic type II MIKC^c MADS-box genes of liverworts, mosses, lycophytes, monilophytes (ferns) and spermatophytes (seed plants). *Cg*MADS1 from *Chara globularis* was used as a representative of the outgroup. Branches leading to genes from the different phyla are coloured according to the simplified phylogeny of land plants that is shown in the top left corner. The phylogenetic position of some known type II MIKC^c MADS-box genes[111] representative of previously described clades of MADS-box genes are indicated on the right together with the species and phylum in which these genes have been identified. The four clades of MIKC^c MADS-box genes that trace back to the most recent common ancestor of Euphyllophytes are shaded in grey.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Assessments of transcriptome assembly gene-family representation relative to gene-family members identified in annotated genomes. a, BUSCO versus CEGMA (CEG) gene occupancy for each sample. BUSCO transcriptome completeness is given as 'complete plus fragmented' BUSCO percentage using the eukaryota_odb9 database. CEGMA transcriptome completeness is given as conditional reciprocal best BLAST hits (see Supplementary Methods). Dotted line represents 57.5% (BUSCO) and 70% (CEGMA) gene occupancy threshold. Black dots represent 1KP samples (n = 1,020) and blue dots annotated plant genomes (n = 30). **b**, BUSCO gene occupancy for each major clade. Boxes represent lower and upper quartiles; the black bold line represents the median and whiskers extend to the most-extreme data points. Sample sizes: Chromista, n = 23; Rhodophyta, n = 18; Glaucophyta, n = 2; Chlorophyta, n = 94; streptophyte algae, n = 42; hornworts, n = 7; liverworts, n = 18; mosses, n = 38; lycophytes, n = 16; monilophytes, n = 59; gymnosperms, n = 76; ANA grade, n = 6; monocots, n = 96; Chloranthales, n = 1; magnoliids, n = 22; CRPT grade, n = 29; asterids, n = 205; Caryophyllales, n = 48; rosids, n = 176; Saxifragales, n = 23; Santalales, n = 6. Dotted line represents 57.5% (BUSCO) gene occupancy threshold. **c**, Scatterplot of gene-family sizes in transcriptomes versus genomes on a logarithmic scale. The grey line indicates $x = y$, the black line indicates a linear regression fitted to the data (n = 299; 23 gene families in 13 species groups). Pearson and Spearman correlation coefficients (n = 299) are indicated. **d**, Box plot of transcriptome:genome ratios of gene-family sizes for each species group. Boxes indicate upper and lower quartiles with median; whiskers extend to data points no more than 1.5× the interquartile range (n = 23) with outliers plotted as individual data points. **e**, **f**, Number of remaining sequences after filtering with cd-hit and a threshold of 100%, 99.9%, 99%, 95% or 90% in transcriptome sequences and reference genomes (Supplementary Table 8). Boxes indicate upper and lower quartiles with median; whiskers extend to data points no more than 1.5× the interquartile range (**e**, n = 1,451; **f**, n = 32) with outliers plotted as individual data points.

nature research

Corresponding author(s): James Leebens-Mack & Gane Ka-Shu Wong

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Transcripts were assembled using SOAPdenovo-Trans assembler (version of 2012-04-05); NCBI BLAST, TransRate, CEGMA6 and BUSCO were used to assess assembly quality, translations were performed using TransPipe and Genewise 2.2.2, Gene and species tree estimates RAxML v. 8.1.17, FastTree-2 v. 2.1.5, and ExaML v. 3.0.14,ASTRAL-II v. 5.0.3 was used to estimate species trees; scripts for post-processing, DiscoVista, of trees - https://github.com/smirarab/1kp ; genome duplications were investigated using the DupPipe, PAML, and the MAPS pipelines including the GuestTreeGen program withinGenPhyloData - https://bitbucket.org/barkerlab/maps ; analysis of gene family expansions included HMMER v3.1b2 and scrips available at https://github.com/GrosseLab/OneKP-gene-family-evo
Data analysis	

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data are public: Raw reads in NCBI SRA database - http://www.onekp.com/public_read_data.html ; Assembled transcripts and transcript translations - http://www.onekp.com/public_data.html ; Gene family nucleotide and amino acid fasta files - http://jlmwiki.plantbio.uga.edu/onekp/v2/ ; Multiple sequence alignments, gene trees and species trees for single copy nuclear genes - https://github.com/smirarab/1kp

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Gene and species phylogenies were estimated in order to infer: relationships across the green tree of life (Viridiplantae), the timing of genome-scale duplication events, and the timing of gene family expansions.
Research sample	RNA was isolated from young vegetative tissue from 1342 samples representing 1147 species across all major subclades of Viridiplantae, glaucophytes (Glaucophyta) and red algae (Rhodophyta) and used to generate RNA seq reads and assemblies.
Sampling strategy	Samples were collected as available in living collections. Species were chosen for RNA seq with a priority to maximize taxonomic diversity across Viridiplantae and outgroups
Data collection	RNA samples were derived from vouchered material in living collections as described in Table 1.
Timing and spatial scale	Samples were collected as available. No attempt was made to control for environmental variation
Data exclusions	RNA samples exhibiting evidence of contamination were excluded from phylogenetic analyses. Contamination was diagnosed through BLAST comparisons to ribosomal RNA and plastid gene databases.
Reproducibility	Bootstrap analyses and Bayesian posterior probabilities were estimated for all nodes in gene trees and species trees.
Randomization	Bootstrap support for nodes gene trees and species trees were estimated in a standard fashion through random resampling of columns in sequence alignments.
Blinding	No blinding was done for any of our analyses.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Most samples are available in live collections and/or herbarium vouchers.