Current Biology

Multiple Metabolic Innovations and Losses Are Associated with Major Transitions in Land Plant Evolution

Highlights

- Angiosperm-like brassinosteroid biosynthetic pathway evolved in a stepwise manner
- Gibberellin biosynthesis and inactivation genes are present in bryophytes
- This suggests bryophytes may contain as-yet-undiscovered bioactive gibberellins
- Known cutin/suberin biosynthetic genes are absent from non-angiosperm plants

Authors

Naomi Cannell, David M. Emms, Alexander J. Hetherington, John MacKay, Steven Kelly, Liam Dolan, Lee J. Sweetlove

Correspondence

lee.sweetlove@plants.ox.ac.uk

In Brief

In a comparative analysis of genes in 72 plants and algae, Cannell et al. identify innovations and losses in metabolic capabilities for biosynthesis/inactivation of phytohormones and biosynthesis of structural polymers over the course of land plant evolution. Analysis of transcriptomes from 305 non-angiosperm plants supports the findings.

Cannell et al., 2020, Current Biology *30*, 1783–1800
May 18, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cub.2020.02.086

CellPress

Current Biology
Article

Multiple Metabolic Innovations and Losses Are Associated with Major Transitions in Land Plant Evolution

Naomi Cannell,[1] David M. Emms,[1] Alexander J. Hetherington,[1] John MacKay,[1] Steven Kelly,[1] Liam Dolan,[1] and Lee J. Sweetlove[1,2,*]
[1]Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK
[2]Lead Contact
*Correspondence: lee.sweetlove@plants.ox.ac.uk
https://doi.org/10.1016/j.cub.2020.02.086

SUMMARY

Investigating the evolution of plant biochemistry is challenging because few metabolites are preserved in fossils and because metabolic networks are difficult to experimentally characterize in diverse extant organisms. We report a comparative computational approach based on whole-genome metabolic pathway databases of eight species representative of major plant lineages, combined with homologous relationships among genes of 72 species from streptophyte algae to angiosperms. We use this genomic approach to identify metabolic gains and losses during land plant evolution. We extended our findings with additional analysis of 305 non-angiosperm plant transcriptomes. Our results revealed that genes encoding the complete biosynthetic pathway for brassinosteroid phytohormones and enzymes for brassinosteroid inactivation are present only in spermatophytes. Genes encoding only part of the biosynthesis pathway are present in ferns and lycophytes, indicating a stepwise evolutionary acquisition of this pathway. Nevertheless, brassinosteroids are ubiquitous in land plants, suggesting that brassinosteroid biosynthetic pathways differ between earlier- and later-diverging lineages. Conversely, genes for gibberellin biosynthesis and inactivation using methyltransferases are found in all land plant lineages. This suggests that bioactive gibberellins might be present in bryophytes, although they have yet to be detected experimentally. We also found that cytochrome P450 oxidases involved in cutin and suberin production are absent in genomes of non-angiosperm plants that nevertheless do contain these biopolymers. Overall, we identified significant differences in crucial metabolic processes between angiosperms and earlier-diverging land plants and resolve details of the evolutionary history of several phytohormone and structural polymer biosynthetic pathways in land plants.

INTRODUCTION

Some time before 450 million years ago, a single lineage of streptophyte algae colonized land, and the subsequent evolutionary radiation gave rise to the embryophytes, which are the dominant photosynthetic organisms in the terrestrial flora [1]. Extant embryophytes display a diverse range of morphologies, physiologies, and biochemistries. The evolution of this diversity had enormous impact on the terrestrial biotic and abiotic environment, affecting nutrient cycles and hydrology and modifying earth sediments and atmosphere [2]. Many of the adaptations that accompanied the colonization of land involved evolution of new metabolic capabilities [3]. Examples include metabolic pathways for the biosynthesis of phytohormones coordinating plant growth, of specialized metabolites that defend against pathogens and provide tolerance of abiotic stresses, and of structural polymers such as lignin and suberin that provide support to stems and roots [4–6]. This diversity of metabolic enzymes is a consequence of substantial adaptive gene family expansion [7].

Investigating the full complement of metabolic pathways in an organism is challenging. Much of our current knowledge of plant metabolism is the result of experimental characterization of the kinetic properties of purified enzymes or *in vivo* metabolic flux analysis using isotope tracers [8]. The laborious nature of these approaches means that we lack a complete picture of metabolism for any given species, even for well-studied angiosperms. Moreover, the focus on a few model species and crops means that current knowledge of plant metabolism is heavily skewed toward angiosperms [9], which means that we have only a partial picture of the evolution of plant metabolism over the course of land plant history.

Recently, *in silico* systems approaches have been applied to plant metabolism [10, 11], facilitating the development of computational representations of entire metabolic networks from genome sequences. Furthermore, the availability of transcriptomes and whole-genome sequences of algae and earlier-diverging land plants has provided researchers with valuable resources for metabolic modeling [12–14]. These data are a source of metabolic information that remains largely uninterrogated. A single study has analyzed the evolution of metabolism from genome sequence data, revealing that the main metabolic innovations after the appearance of vascular plants relate to specialized metabolism [15]. However, the study analyzed just 16 species providing a sparse sample of plant evolution and only provided a top-level overview of metabolic pathway types, not specific functions.

Here, we exploit a wider range of genome sequence and transcriptome data to investigate in detail the evolution of metabolism in the Chloroplastida. A comparative approach was used to identify metabolic innovations and losses occurring during the evolution of streptophyte algae and land plants. Our starting point was the generation of genome-scale metabolic pathway databases for seven species spanning from algae to land plants (two charophytes, two bryophytes, a lycophyte, a monilophyte, and a gymnosperm). From these databases, and an additional pre-existing database for *Arabidopsis* [16], the presence and absence of the full range of known metabolic pathways (and therefore the appearance and disappearance of various metabolic traits over the course of land plant evolution in these species) was inferred. Individual metabolic gene annotations from these databases were then used in combination with information on homologous genes in other organisms to confirm inferred metabolic innovations or losses and extend their supporting evidence across 64 further plant and algal genomes (bringing the total number of species analyzed to 72). To provide further support for metabolic innovations, an additional 305 non-angiosperm transcriptomes were analyzed. The identification of several known evolutionary innovations in plant metabolism validates the method. The analysis revealed new information on the occurrence of biosynthetic enzymes for phytohormones and structural compounds across land plants and algae.

RESULTS

Bioinformatic Approach and Justification

The metabolic capabilities of an organism can be inferred from its genome sequence. However, even when the genome is well annotated, this is a time-consuming process, involving complex algorithms to infer the presence or absence of whole metabolic pathways. Therefore, rather than attempting to make a detailed comparison of metabolic pathways present in all plant species for which genome sequences are available, the initial comparison was simplified to a smaller subset of species representing each of the major lineages in plant evolution, i.e., charophyte algae, bryophytes, lycophytes, ferns, gymnosperms, and angiosperms. The choice of these organisms was made to maximize the breadth of plant phylogeny covered while maintaining a feasible number of datasets for analysis. Since well-curated information on *Arabidopsis* metabolism is already available [16], efforts here were focused on earlier-diverging land plants and algae and directed mainly based on the availability of sequenced genomes, which are sparse for such organisms. In fact, sequence data for four of the included species have only become available in the last 2 years. Two streptophyte algae genomes have been sequenced. Both—*Klebsormidium nitens* [13] and *Chara braunii* [17]—were included in this analysis. Two bryophytes were chosen: the moss *Physcomitrella patens* [12], a model organism and the bryophyte with the most well-developed genome annotation, and *Marchantia polymorpha* [14], an emerging model organism. *Selaginella moellendorffii* [18] was selected as a representative lycophyte and the fern *Salvinia cucullata* [19] was selected as a representative monilophyte. Finally, *Picea glauca* was selected as a representative gymnosperm [20]. Note that due to concerns about the reliability of assembly of the conifer megagenome, we used transcriptome data for this species.

Metabolic pathways from the MetaCyc [21] and PlantCyc [22] databases were identified from the genome sequences of these seven species using the Pathway Tools software [23]. To increase the robustness of the analysis, two parallel Pathway Tools analyses were carried out, one based on annotation files compiled from each species' published genome annotation, and one using EC number annotations generated from E2P2, a machine learning-based algorithm for metabolic annotation [22]. The union of the predicted sets of pathways from these two approaches formed the final pathway/genome database (PGDB) for each species. Possible metabolic innovations and losses were inferred when the number of genes encoding enzymes associated with metabolic pathways either increased or decreased at any point, suggesting the gain or loss of the metabolic pathway at the branchpoint in question.

Because the eight species are only a representative sample of a large phylogenetic space, each identified metabolic gain or loss across the eight species was mapped onto genetic data for 72 phylogenetically ordered species from the Chloroplastida (a full, labeled phylogeny of all included species is provided as Figure S1). The species investigated consist of 46 angiosperms, seven gymnosperms (*Ginkgo biloba* [24], *Gnetum montanum* [25], *Pseudotsuga menziesii* [26], *Pinus lambertiana* [27], *Pinus taeda* [28], *Picea glauca* [20], *Picea abies* [29]), two ferns (*Salvinia cucullata* [19], *Azolla filiculoides* [19]), three lycophytes (*Selaginella moellendorffii* [18], *Selaginella tamariscina* [30], *Isoetes echinospora* (unpublished data; GenBank: GGKY00000000.1), two mosses (*Physcomitrella patens* [12], *Sphagnum fallax* [31]), one liverwort (*Marchantia polymorpyha* [14]), four charophytes (*Klebsormidium nitens* [13], *Chara braunii* [17], *Spirogloea muscicola* [32], *Mesotaenium endlicherianum* [32]), and seven chlorophytes (*Ostreococcus lucimarinus*, *Micromonas sp. RCC299*, *Micromonas pusilla CCMP1545*, *Coccomyxa subellipsoidea C-169*, *Chromochloris zofingiensis*, *Volvox carteri*, *Chlamydomonas reinhardtii* [31]). For 70 of these species, genome sequences were used; for *Picea glauca* and *Isoetes echinospora*, transcriptome data were used. All genes from the 72 species were placed into orthogroups using OrthoFinder software [33]. Each of the metabolic gains/losses identified from the eight-species comparison was then tracked across the 72-species orthogroups using gene trees. In the following sections, only metabolic innovations and losses that held true beyond the eight representatives of the major evolutionary groups are presented. To provide additional support for results that relate to non-angiosperms (which are under-represented in terms of genome availability), 305 non-angiosperm land plant and algal transcriptomes taken from the One Thousand Plant Transcriptome Initiative [34] were analyzed to identify homologs of the genes required for the presented metabolic pathways in these species.

De Novo Genome Annotation Provides Gene-to-Reaction Associations for Previously Unannotated Metabolic Genes

EC number annotations identified using E2P2 were compared to published genome annotations for each of the seven non-angiosperm land plants and algae for which PGDBs were generated. Novel metabolic gene annotations from E2P2 were identified in all seven species, both in the form of completely new annotations for genes previously described only as open reading frames and as additional annotations for genes already associated with another function or metabolic reaction. An overview of the number and classification (i.e., completely new annotations or additional annotations) of novel metabolic gene annotations for each species is shown in Figure S2, and tables containing EC number and MetaCyc reaction ID annotations produced by E2P2 for each species are provided in Data S1, which also contains full PGDBs for each species based on the union of both annotations, provided as lists of MetaCyc/PlantCyc metabolic pathway IDs. Additionally, orthogroups identified using OrthoFinder across the complete set of 72 species is provided in Data S2, and the gene trees for each orthogroup are provided in Data S3. A Python script allowing this information to be queried based on associations between MetaCyc/PlantCyc reaction identifiers and *Arabidopsis* genes is supplied as Data S4.

Inferred Metabolic Pathways Capture Known Phylogenetic Relationships

To confirm that the PGDBs of representative species contain information applicable to the study of plant evolution, k-mediods clustering followed by a dimensionality reduction technique known as t-SNE [35] was carried out to generate a cluster map of metabolic pathway similarity among the species to examine how this relates to phylogeny. The analysis grouped species based on the presence and absence of the metabolic pathways in each species' PGDB. As can be seen in Figure S3, the green algae form a two-member cluster distinct from the land plants. Terrestrial plants were split into two clusters, one containing all land plants up to and including *Picea* and the other containing only *Arabidopsis* (Figure S3). This isolation of *Arabidopsis* is potentially a consequence of the presence of angiosperm-specific metabolic pathways in its PGDB. However, it may also be that the difference is influenced by the level of manual curation involved in the production of the *Arabidopsis* PGDB—the other PGDBs developed in this analysis were not manually curated. Nevertheless, the nearest neighbors of *Arabidopsis* as placed by t-SNE on this graph are the two next most recently diverging land plants, the euphyllophytes *Picea* and *Salvinia*. The two bryophytes, *Physcomitrella* and *Marchantia*, clustered together and were grouped with the earliest diverging member of the vascular plants, the lycophyte *Selaginella.*

Overall, a clear distinction between earlier-diverging and later-diverging lineages from streptophyte algae to angiosperms can be seen on the graph. Based only on the presence and absence of metabolic pathways in each species' PGDB, it was thus possible to capture the known phylogenetic relationships between the plant species analyzed (Figure S3). This indicates that the metabolic information gathered and included in the PGDBs reflects true differences between the species and supports the use of such information in the analysis of the evolution of metabolism.

Comparative Analysis Identifies Previously Identified Metabolic Innovations and Losses in Land Plant Evolution

Orthogroup analysis identified 50 metabolic pathways that showed gains or losses over the course of land plant evolution (Figure 1). To prioritize metabolic pathways for further investigation, we considered the novelty of the observation and the number of reactions in the pathway. The latter criterion was used because some of the metabolic "pathways" as categorized by Pathway Tools are extremely short, containing only one or two reactions and these were not considered further. The 17 pathways highlighted by dark blue circles in Figure 1 were the ones that were further investigated. Table S1 contains orthogroup and gene IDs for each of the pathways investigated in the following sections.

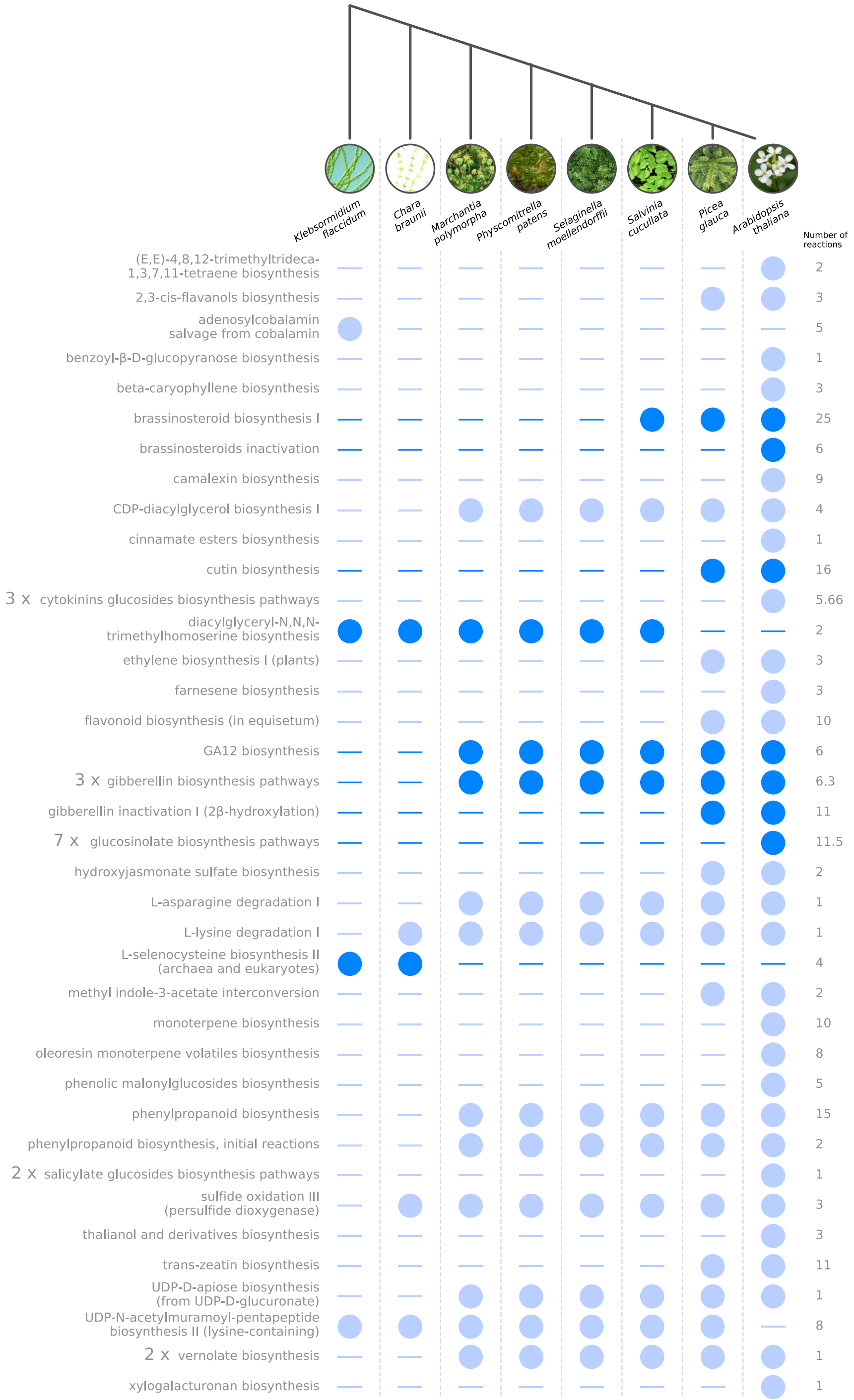
The reliability of the approach was demonstrated by the fact that three known metabolic changes during land plant evolution were correctly identified. The first example is the metabolic pathway for diacylglyceryl-N,N,N-trimethylhomoserine (DGTS) biosynthesis, which is thought to have been lost in spermatophytes where DGTS has been replaced by phosphatidylcholine [36–38]. Our analysis picked out this metabolic pathway because of the presence of genes encoding the required enzyme only in the genomes of algae, bryophytes, lycophytes, and ferns (Figures 1 and S4). Homologous genes were not identified in any gymnosperm or angiosperm species. This pattern is consistent with the previous identification of DGTS in ferns but not seed plants [36, 37].

The second example is the capacity for biosynthesis of selenocysteine, which is considered an ancestral character, found in bacteria, mammals, and green algae but lost in land plants [39]. The pathway of selenocysteine biosynthesis was identified by our approach as a metabolic loss during land plant evolution (Figure 1). The gene encoding the first enzyme in the pathway is found in all plant taxa analyzed, while the genes encoding the full combination of enzymes required for this pathway are found only in the algae (Figure S5).

Finally, our analysis identified several metabolic pathways involved in glucosinolate biosynthesis from various amino acids as a metabolic capability that has been gained during land plant evolution (Figure 1). We found only one or two genes encoding the involved enzymes in algae and non-angiosperm plants and never the complete pathway (Figure 2B). The genes were more commonly found in angiosperms, but the complete pathway was only encoded by the genomes of the Brassicaceae in our analysis (Figure 2). This is in agreement with the hypothesis that glucosinolates are a synapomorphy of the Capparales [40–42]. The analysis shown in Figure 2 also reveals a more detailed picture of the evolution of glucosinolate biosynthesis. The presence of genes for several different glucosinolate-related enzymes across the angiosperms is unsurprising, given that several of these encoding genes are members of large gene families (e.g., glucosyltransferases and monooxygenases) and the evolution of glucosinolate-specific enzymes likely evolved from similar pre-existing pathways such as cyanogenic glucosides production [43]. However, the presence of genes encoding the enzyme glucosinolate γ-glutamyl peptidase (GGP) in gymnosperms and flavin-containing monooxygenases (GS-OX) in species of every clade from chlorophyte algae to angiosperms suggests a conserved ancient function for these genes outside of the biosynthesis of glucosinolates.

Biosynthetic Capabilities for Gibberellin and Brassinosteroid Production Are Encoded in the Genomes of Non-angiosperm Plants

The appearance of enzymatic capabilities for synthesis of certain plant hormones has been linked to evolutionary adaptations to



	Klebsormidium flaccidum	Chara braunii	Marchantia polymorpha	Physcomitrella patens	Selaginella moellendorffii	Salvinia cucullata	Picea glauca	Arabidopsis thaliana	Number of reactions
(E,E)-4,8,12-trimethyltrideca-1,3,7,11-tetraene biosynthesis	—	—	—	—	—	—	—	○	2
2,3-cis-flavanols biosynthesis	—	—	—	—	—	—	○	○	3
adenosylcobalamin salvage from cobalamin	○	—	—	—	—	—	—	—	5
benzoyl-β-D-glucopyranose biosynthesis	—	—	—	—	—	—	—	○	1
beta-caryophyllene biosynthesis	—	—	—	—	—	—	—	○	3
brassinosteroid biosynthesis I	—	—	—	—	—	●	●	●	25
brassinosteroids inactivation	—	—	—	—	—	—	—	●	6
camalexin biosynthesis	—	—	—	—	—	—	—	○	9
CDP-diacylglycerol biosynthesis I	—	—	○	○	○	○	○	○	4
cinnamate esters biosynthesis	—	—	—	—	—	—	—	○	1
cutin biosynthesis	—	—	—	—	—	—	●	●	16
3 x cytokinins glucosides biosynthesis pathways	—	—	—	—	—	—	—	○	5.66
diacylglyceryl-N,N,N-trimethylhomoserine biosynthesis	●	●	●	●	●	●	—	—	2
ethylene biosynthesis I (plants)	—	—	—	—	—	—	○	○	3
farnesene biosynthesis	—	—	—	—	—	—	—	○	3
flavonoid biosynthesis (in equisetum)	—	—	—	—	—	—	○	○	10
GA12 biosynthesis	—	—	●	●	●	●	●	●	6
3 x gibberellin biosynthesis pathways	—	—	●	●	●	●	●	●	6.3
gibberellin inactivation I (2β-hydroxylation)	—	—	—	—	—	—	●	●	11
7 x glucosinolate biosynthesis pathways	—	—	—	—	—	—	—	●	11.5
hydroxyjasmonate sulfate biosynthesis	—	—	—	—	—	—	○	○	2
L-asparagine degradation I	—	—	○	○	○	○	○	○	1
L-lysine degradation I	—	○	○	○	○	○	○	○	1
L-selenocysteine biosynthesis II (archaea and eukaryotes)	●	●	—	—	—	—	—	—	4
methyl indole-3-acetate interconversion	—	—	—	—	—	—	○	○	2
monoterpene biosynthesis	—	—	—	—	—	—	—	○	10
oleoresin monoterpene volatiles biosynthesis	—	—	—	—	—	—	—	○	8
phenolic malonylglucosides biosynthesis	—	—	—	—	—	—	—	○	5
phenylpropanoid biosynthesis	—	—	○	○	○	○	○	○	15
phenylpropanoid biosynthesis, initial reactions	—	—	○	○	○	○	○	○	2
2 x salicylate glucosides biosynthesis pathways	—	—	—	—	—	—	—	○	1
sulfide oxidation III (persulfide dioxygenase)	—	○	○	○	○	○	○	○	3
thalianol and derivatives biosynthesis	—	—	—	—	—	—	—	○	3
trans-zeatin biosynthesis	—	—	—	—	—	—	○	○	11
UDP-D-apiose biosynthesis (from UDP-D-glucuronate)	—	—	○	○	○	○	○	○	1
UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing)	○	○	○	○	○	○	○	—	8
2 x vernolate biosynthesis	—	—	○	○	○	○	○	○	1
xylogalacturonan biosynthesis	—	—	—	—	—	—	—	○	1

Figure 1. Overview of Metabolic Gains and Losses Identified by the Orthogroup Analysis

For each representative species, the presence or absence of the metabolic pathway in question is denoted by a filled circle (presence) or a horizontal line (absence). The far-right column shows the number of reactions in each metabolic pathway; where multiple similar reactions are grouped, e.g., for glucosinolate biosynthesis, this number denotes the average number of reactions across the grouped pathways. Dark blue circles indicate the metabolic pathways discussed in further detail. All metabolic pathway names are taken from the MetaCyc or PlantCyc metabolic pathway databases; Table S2 contains a list of the MetaCyc/PlantCyc IDs corresponding to each of the metabolic pathway names listed. Figure S1 shows a full phylogeny for the 72 species across which these metabolic gains/losses were identified, while Figure S2 contains an overview of the additional EC number annotations for each representative species for which PGDBs were constructed. Figure S3 shows the results of clustering these PGDBs based on metabolic pathway content.

terrestrial environments [4, 44]. Consistent with this, genes encoding enzymes involved in both gibberellin (GA) and brassinosteroid biosynthesis were associated with the transition to land in the 72-species comparison—i.e., were absent from all algal species analyzed (Figures 3 and 4). The enzymes comprising the GA biosynthesis pathway from *ent*-kaurene to GA_{12} and its conversion to bioactive forms (Figure 3A) were not found in the charophyte or chlorophyte algal genomes (Figure 3B). Similarly, genes encoding enzymes for the majority of the metabolic pathway of brassinosteroid biosynthesis were not found in these two algal groups (Figure 4B). The algae do, however, contain homologs of DET2 encoding α5 steroid dehydrogenase, the first enzyme in the pathway that our analysis suggests is present throughout the land plant lineage (Figure 4B).

Previously, it has been suggested that GAs are confined to vascular plants [45]. However, our analysis demonstrates that the enzymes required for their biosynthesis are encoded in bryophyte genomes. The two enzymes responsible for biosynthesis of the inactive GA precursor GA_{12} (*ent*-kaurene oxidase [KO] and kaurenoic acid oxidase [KAO] [Figure 3A]) are present in all land plants except the mosses *Physcomitrella patens* and *Sphagnum fallax*, which are missing KAO (Figure 3B). This supports previous reports that *P. patens* is missing KAO [7, 46, 47] and extends this finding to an additional moss species. The presence of KO in all land plants is consistent with observations that its metabolite product, the GA_{12} intermediate KA, has been widely found in plants [48, 49]. However, there has been no metabolomic evidence for the occurrence of recognized active forms of GA in bryophytes [47, 50]. Yet, our analysis provides evidence for the presence of genes encoding GA20ox and GA3ox enzymes in all land plant clades analyzed, including the liverworts. These enzymes are 2-oxoglutarate-dependent dioxygenases responsible for the oxidation of GA_{12} into bioactive GAs (Figure 3A). GA oxidase homologs have been previously identified in *P. patens* [51], and the identification here of homologs in

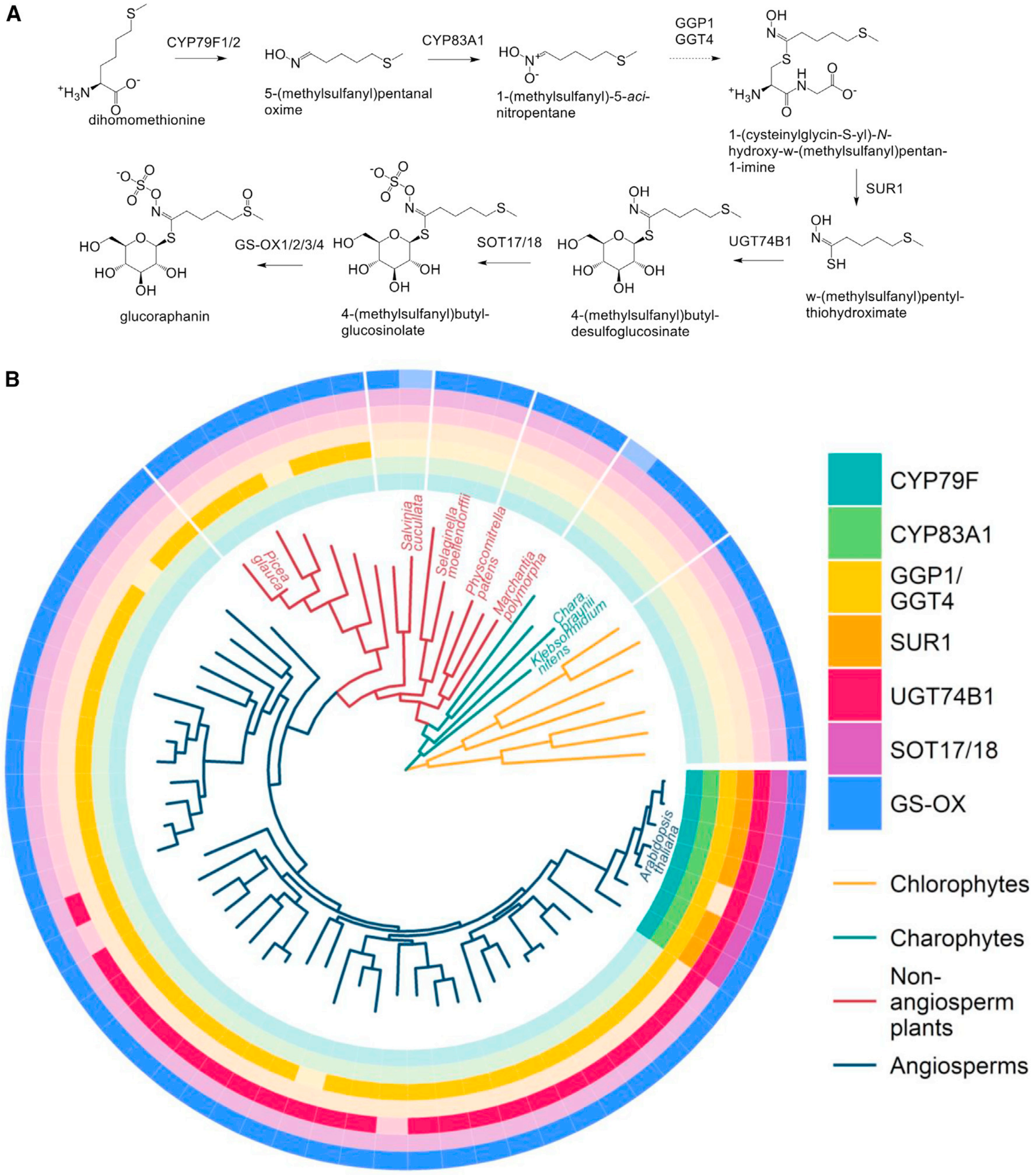


Figure 2. Presence of Gene Homologs Encoding Glucosinolate Biosynthesis Enzymes across 72 Plant and Algal Species

(A) Biosynthetic pathway for glucosinolates. Although several variations on this pathway were identified as possible evolutionary transitions (all utilizing the same enzymes), only glucosinolate biosynthesis from a dihomomethionine precursor is shown.

(B) Homolog presence (colored rings) for genes involved in glucosinolate biosynthesis across 72 studied species. Dark colors indicate the presence of a homolog while light colors indicate absence. Orthogroup information for the genes involved is provided in Table S1. Figures S4 and S5 show the presence of gene homologs for the additional validation pathways - DGTS and selenocysteine biosynthesis. CYP, cytochrome P450; GGP, γ-glutamyl peptidase; GGT, γ-glutamyl transpeptidase; SUR, alkyl-thiohydroximate C-S lyase; UGT, UDP-glucosyltransferase; GS-OX, glucosinolate S-oxygenase.

moss *Sphagnum fallax* and the liverwort *Marchantia polymorpha* supports the evolution of GA oxidases in bryophytes.

Conversely, beyond the initial reaction catalyzed by enzymes encoding DET2, our analysis did not identify genes encoding brassinosteroid biosynthesis enzymes in bryophytes or algae. Following the synthesis of campestanol by the ubiquitous DET2 enzyme, brassinosteroid biosynthesis follows a series of oxidative modifications in two overlapping pathways (Figure 4A) catalyzed by cytochrome P450 (CYP) enzymes. All *Arabidopsis* enzymes involved are members of the CYP85 clan—either the CYP85A subfamily (involved in C6 oxidation and ring extension) or the CYP90A/B/C/D subfamilies (which catalyze C22 and C23

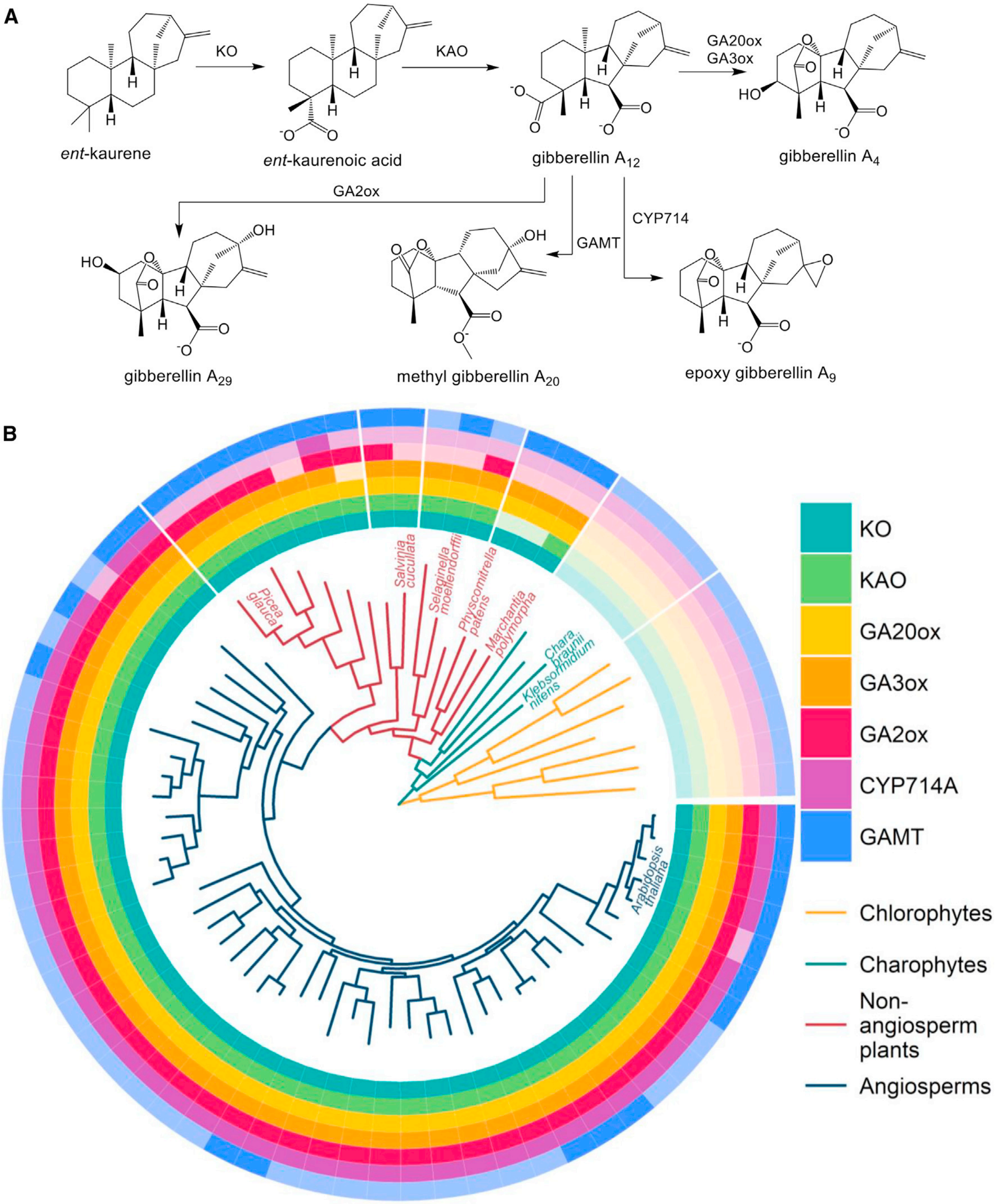


Figure 3. Presence of Gene Homologs Encoding Gibberellin Biosynthesis and Inactivation Enzymes across 72 Plant and Algal Species

(A) The gibberellin biosynthetic/inactivation pathway.

(B) Phylogeny showing the presence and absence of homologs for genes (colored rings) involved in gibberellin biosynthesis across 72 analyzed species. Darker ring colors indicate gene presence while lighter colors indicate absence. Internal phylogeny colors show broad plant groupings. Orthogroup information for the genes involved is provided in Table S1. KO, *ent*-kaurene oxidase; KAO, *ent*-kaurenic acid oxidase; GA20ox, gibberellin 20 oxidase; GA3ox, gibberellin 3 oxidase; GA2ox, gibberellin 2 oxidase; CYP, cytochrome P450; GAMT, gibberellin methyltransferase.

hydroxylation). The brassinosteroid castasterone has been identified in bryophytes, albeit at concentrations that are orders of magnitude lower than in angiosperms [52]. The lack of CYP85A and CYP90 homologs in bryophytes and algae suggests either that the biosynthetic route for castasterone diverges from that shown in Figure 4A in these species or that other cytochrome P450s are involved in its biosynthesis. Green algae are known to contain CYP85 clan enzymes with divergent sequences from

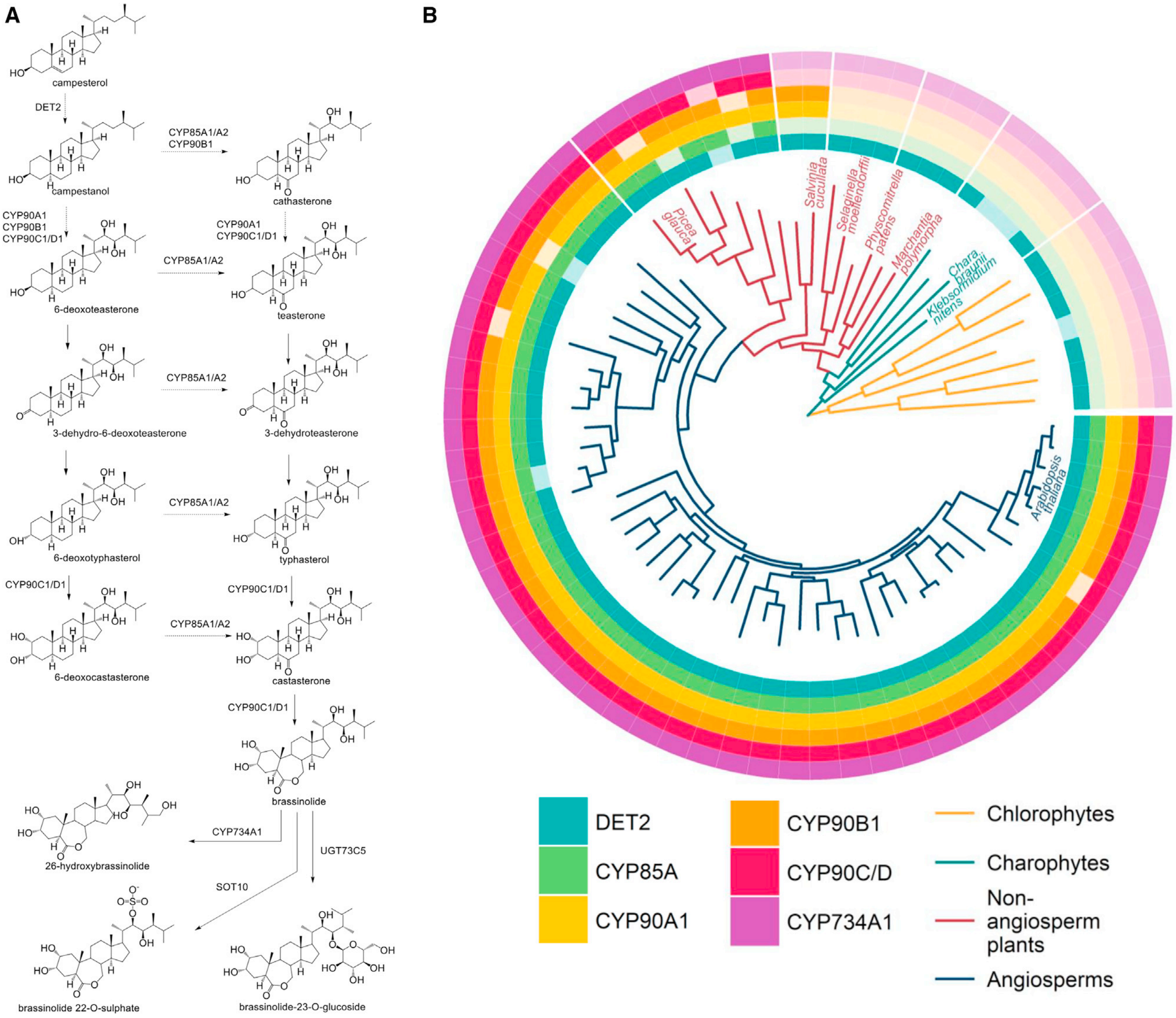


Figure 4. Presence of Gene Homologs Encoding Brassinosteroid Biosynthesis and Inactivation Enzymes across 72 Plant and Algal Species

(A) The brassinosteroid biosynthetic/inactivation pathway.

(B) Evidence (colored rings) for homologs of brassinosteroid biosynthetic enzymes across 72 plant species. Darker colors indicate presence of gene homologs, while lighter colors indicate absence. Internal phylogeny colors show broad plant groupings. Degradation steps are shown only for brassinolide; however, the same enzymes also act to degrade castasterone. Orthogroup information for the genes involved is provided in Table S1. A simplified gene tree showing relationships between homologous genes across species in the orthogroup containing CYP734A1 is provided as Figure S6A. DET2, 5α steroid dehydrogenase; CYP, cytochrome P450; UGT, UDP glucose:cytokinin glucosyltransferase; SOT, brassinosteroid sulfotransferase.

those in angiosperms [7]—it is possible that these genes and their relatives in the bryophytes are responsible for brassinosteroid biosynthesis in these species but were replaced by CYP85A and CYP90 genes in spermatophytes. The latter scenario may also explain our finding of the absence of homologs of genes encoding brassinosteroid biosynthesis enzymes in the lycophytes (Figure 4B). Brassinosteroids have been found in lycophytes [52, 53], but we did not find specifically identifiable CYP85A and CYP90A/B/C/D homologs in either the *S. moellendorffi* and *S. tamariscina* genomes or in the *I. echinospora* transcriptome (Figure 4B). However, we did identify genes in *S. moellendorffii* and *S. tamariscina* forming a sister group to the clade containing specific CYP90C and CYP90D genes in other species. While these genes are not specifically identifiable as CYP90 class C or D, their close relationship suggests they may be candidate genes for brassinosteroid biosynthesis in lycophytes. Six gymnosperm genomes were found to encode specific CYP90D1 genes. CYP90C1 is the closest relative of CYP90D1 in angiosperms; presumably, the former arose from a gene duplication after the divergence of the gymnosperms. However, these genes are functionally redundant [54], and we therefore consider the angiosperm-like brassinosteroid biosynthesis pathway to be complete in the gymnosperms. Three of the gymnosperm genomes analyzed contain homologs of all angiosperm brassinosteroid biosynthetic genes, while the genomes of the ferns *S. cucullata* and *A. filiculoides* contain three biosynthetic genes (and are missing CYP90C/D and CYP85A homologs; Figure 4B), suggesting that brassinosteroid biosynthesis in ferns represents an intermediate stage of evolution before the complete angiosperm-like pathway was established in the gymnosperms.

Non-angiosperm Plants Lack Genes Encoding Known Enzymes for GA and Brassinosteroid Inactivation

A key feature of the use of specialized metabolites as hormones in angiosperms is rapid regulation of the concentration of bioactive forms of the molecules by enzymatic conversion to inactive forms. Our analysis suggests that many of these mechanisms are absent from non-angiosperm plants (Figures 3 and 4). GAs are inactivated in angiosperms using either GA 2-oxidases, GA 13-oxidases, GA 16,17-oxidases, or GA methyltransferases (GAMTs) [55, 56]. Two orthogroups were identified that contained genes encoding GA 2-oxidases, one containing genes encoding enzymes that act on C19 GAs and one containing those that act on C20 GAs. With the exception of single homologs in *Isoetes echinospora*, *Azolla filiculoides*, and *Ginkgo biloba*, GA 2-oxidases acting on C20 GAs were not found in non-angiosperm plants. C19-acting homologs were not found in non-seed plants but were found in six gymnosperm species, while every angiosperm species contained both C19 and C20-acting homologs (Figure 3B).

Cytochrome P450 enzymes encoding GA 13-oxidases and GA 16,17-oxidases have been identified in rice (CYP714B1, CYP714B2, CYP714D1) and *Arabidopsis* (CYP714A1, CYP714A2) [57–59]. These genes are all members of a single orthogroup containing genes exclusively from angiosperm species (Figure 3B) and *G. biloba*. One explanation for the presence of a CYP714A homolog in *G. biloba* is that GA inactivation via CYP714A enzymes may have evolved before the divergence of the gymnosperms but was lost in gymnosperms other than *G. biloba*. An additional GA 13-oxidase (CYP72A9) has recently been identified in *Arabidopsis* [56]. CYP72A9 is the only CYP72A gene that encodes an enzyme with GA13ox activity in *Arabidopsis*. Our orthogroup analysis shows that similar CYP72A genes are present in other dicots—but the close grouping of CYP72A9 with other CYP genes in *Arabidopsis* makes it difficult to suggest candidate genes with GA 13-oxidase activity in other species.

Arabidopsis GAMTs are members of the SABATH family of methyltransferases, which act on various plant hormones and signaling molecules [60]. As shown in Figure 3B, homologs of *Arabidopsis* GAMT1 and GAMT2 are inconsistently distributed across land plants and are not found in algae. The pattern of GAMT homology could mean that genes in several branches of earlier-diverging land plants have convergently evolved methyltransferases with similar sequences to *Arabidopsis* GAMTs. Or, more likely, that GAMT-like methyltransferases are ancestral, and there has been sequence divergence in most angiosperms and some lycophytes. Regardless, it appears that bryophytes, lycophytes, ferns, and gymnosperms have the capacity to inactivate GAs, despite GAs having not yet been found in bryophytes.

There are several mechanisms for brassinosteroid inactivation in angiosperms including hydroxylation, glucosylation, and sulfonation [61–63]. Each of these mechanisms is catalyzed by a single gene product in *Arabidopsis*. The known *Arabidopsis* brassinosteroid hydroxylation enzyme is CYP734A1, also known as CYP72B1. This is a member of a large orthogroup containing CYP72 clan genes including the CYP714 family responsible for GA inactivation. Although this orthogroup includes algal homologs, gene trees place the algal genes as an outgroup, suggesting that they are homologous to the ancestor of all angiosperm CYP genes in this orthogroup. Bryophyte, lycophyte and fern genes in this orthogroup cluster with alternative CYP genes, and there are no direct homologs of CYP734A1 in these species (Figure S6A).

Glucosylation of brassinosteroids in *Arabidopsis* is carried out by a UDP glucosyltransferase enzyme, UGT73C5, while sulfonation is carried out by brassinosteroid sulfotransferase enzyme (SOT10). Orthogroup analysis places both these genes as members of large orthogroups containing other glucosyltransferases and sulfotransferases, respectively. However, both *Arabidopsis* genes in question were grouped closely with additional *Arabidopsis* genes known to have functions outside of brassinosteroid inactivation—we were therefore unable to distinguish homologs in other species.

Bryophytes, Lycophytes, Ferns, and Gymnosperms Lack Specific Cytochrome P450 Family Enzymes Involved in Oxidation of Structural Monomers

The transition of plants to terrestrial environments was accompanied by the evolution of mechanisms that protect against desiccation, UV radiation, and exposure to a new complement of pathogens [64, 65]. The ability of land plants to withstand these stresses results in part from the evolution of several amino acid/lipid-based insoluble biopolymers such as cutin, suberin, and sporopollenin. Our analysis identified the cutin biosynthetic pathway as a metabolic innovation during land plant evolution. As all three of these structural polymers are produced via biosynthetic pathways that use very similar (or overlapping) biochemical reactions, and several orthogroups are shared between pathways, all three pathways were investigated. To date, suberin has been found only in vascular plants, whereas cutin has been found across the land plants. Sporopollenin is found across land plants and has been isolated from the cell walls of several species of chlorophyte and charophyte algae [64, 66]. All three polymers are formed primarily of fatty-acid-derived monomers of varying chain lengths (sporopollenin C12–C18, cutin C16–C18, and suberin C18–C24 fatty acids), and suberin also contains significant amounts of phenolic compounds [67, 68]. The metabolic pathways for all three polymers rely on the activation of free fatty acids with coenzyme A (CoA) and use cytochrome P450 enzymes for the oxidation of fatty acids into polymer-specific monomers, several of which are shared by cutin and suberin (Figure 5A).

We identified homologs of the LACS genes required for fatty acid activation in cutin and suberin biosynthesis in the genomes of sampled land plants from all clades as well as in charophyte and chlorophyte algal genomes (Figure 5B). For sporopollenin biosynthesis, homologs of the ACOS5 gene responsible for the same process were not found in algae but were present in all land plants, while the ACH enzymes that remove CoA were found in charophyte algae and land plants (Figure 6B).

This suggests that the capability for biosynthesis of precursor molecules required for the production of cutin/suberin monomers were present prior to the evolution of land plants. However,

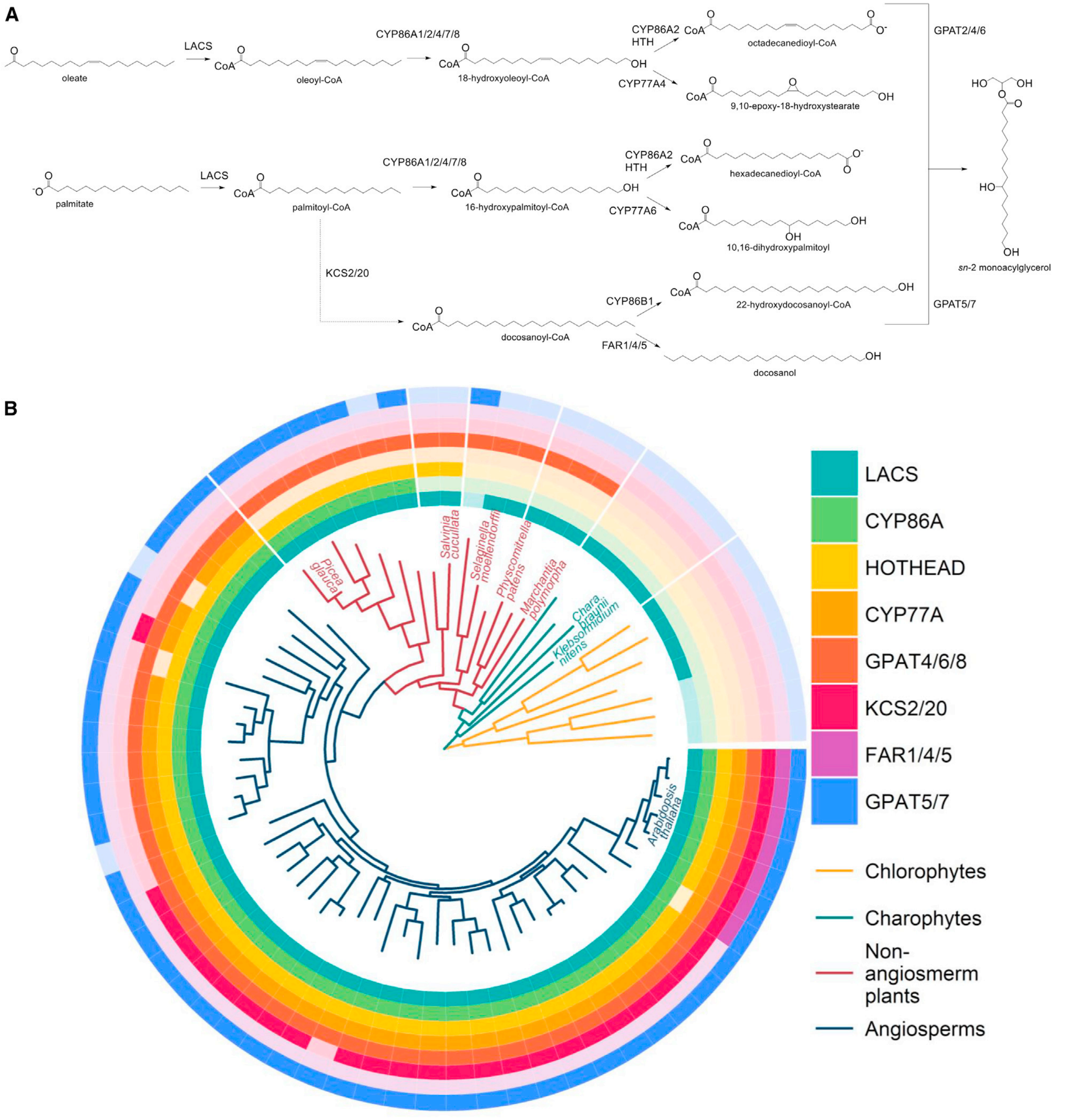


Figure 5. Presence of Gene Homologs Encoding Cutin and Suberin Biosynthesis Enzymes across 72 Plant and Algal Species

(A) The cutin/suberin biosynthetic pathway.

(B) Evidence (colored rings) for homologs of required enzymes across 72 plant species. Darker colors indicate presence of gene homologs while lighter colors indicate absence. Internal phylogeny colors show broad plant groupings. Orthogroup information for the genes involved is provided in Table S1. A simplified gene tree showing relationships between homologous genes across species in the orthogroup containing CYP86 genes is provided as Figure S6B. LACS, long-chain acyl-CoA synthetase; CYP, cytochrome P450; KCS, 3-ketoacyl CoA synthase; FAR, fatty acid reductase; GPAT, glycerol-3-phosphate acyltransferase.

the majority of homologs of the *Arabidopsis* genes encoding enzymes responsible for subsequent reactions in both metabolic pathways were identified only in land plants (Figures 5B and 6B)—this is consistent with the lack of cutin and suberin in algal species [68]. For sporopollenin, the lack of ACOS5 homologs in algae that do nevertheless contain sporopollenin suggests the use of a different, as-yet-unknown, metabolic pathway for sporopollenin biosynthesis in chlorophytes and charophytes.

Further investigation of the gene tree of the orthogroup containing *Arabidopsis* CYP86A and CYP86B subfamily genes,

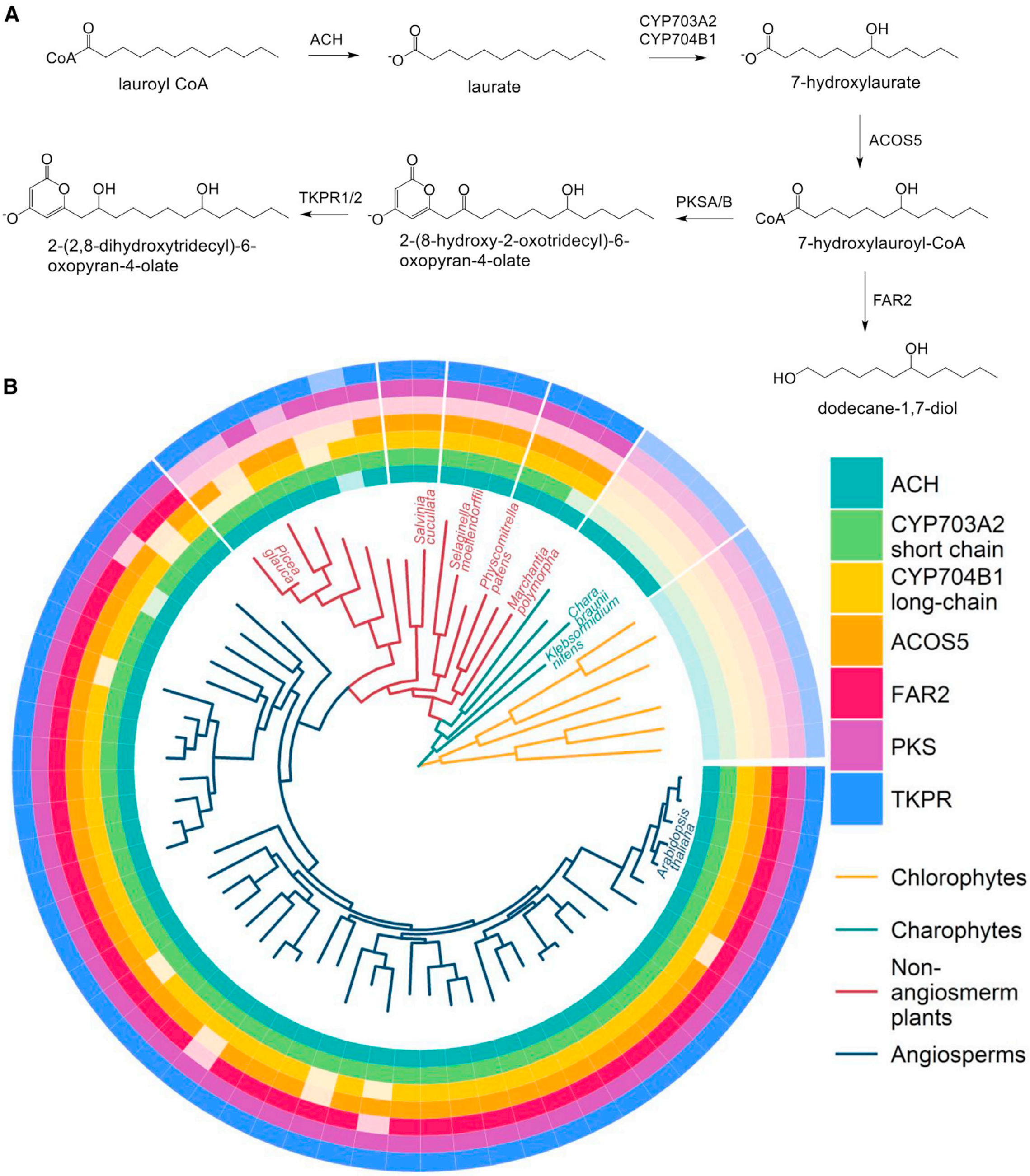


Figure 6. Presence of Gene Homologs Encoding Sporopollenin Biosynthesis Enzymes across 72 Plant and Algal Species

(A) The sporopollenin biosynthesis pathway. Although sporopollenin is produced via analogous reactions from several fatty acid precursors, the depicted pathway shows only those with a lauric acid precursor.

(B) Phylogeny of 72 species showing the presence and absence of homologs of sporopollenin biosynthetic genes (colored rings). Darker colors indicate the presence of homologs while lighter colors indicate absence. Internal phylogeny colors show broad plant groupings. Orthogroup information for the genes involved is provided in Table S1. ACH, acyl-CoA thioesterase; CYP, cytochrome P450; ACOS5, fatty-acyl CoA synthetase; FAR, fatty acid reductase; PKS, polyketide synthase; TKPR, tetraketide pyrone reductase.

responsible for the ω-hydroxylation of fatty acids in cutin and suberin biosynthesis, shows that, while gymnosperm and angiosperm species contain specific homologs of these *Arabidopsis* genes (Figure 5B), earlier-diverging land plants and algae contain homologs that can only be identified more broadly as related to the CYP86 and CYP96 subfamilies (Figure S6B). Direct homologs of the oxidoreductase gene HOTHEAD are also identified in the spermatophytes as well as in both fern species investigated but not in the lycophytes, bryophtyes, or algae (Figure 5B). These gene products in earlier-diverging land plants may also catalyze the ω-hydroxylation of fatty acids but could differ in characteristics such as chain-length preferences. This would be consistent with previously described shorter chain -lengths in the cutin biopolymers of earlier-diverging land plants [69]. However, the lack of direct subfamily homologs mean that no specific function can be inferred for these species.

Similarly, we found CYP77A homologs, involved in in-chain hydroxylation and/or epoxidation, across angiosperm species. However, related genes identified in other land plants could only be classified more generally as CYP77 genes (Figure 5B). We found no CYP77 homologs in algae (Figure 5B), which is as expected given the lack of evidence for the occurrence of cutin or suberin in algae [70]. The presence of genes encoding the LACS enzyme across land plants, combined with the lack of specific CYP86A and CYP77A genes in non-seed plants and the lack of HOTHEAD in lycophytes and bryophytes, suggests that the early stages of cutin and suberin biosynthesis are conserved. However, the fatty acid monomers of cutin and suberin must be biosynthesized via alternative gene products (likely from CYP86/77 families) in bryophytes and lycophytes and later-diverging land plants, which may conceivably produce monomers of different chain lengths and with different molecular modifications. The evolution of HOTHEAD before the divergence of ferns, CYP86A/B in the ancestor of spermatophytes and CYP77A only in the angiosperms suggests that the angiosperm-like pathway for cutin and suberin biosynthesis may have evolved in stages, replacing an alternative set of enzymes in the bryophytes and lycophytes.

Regarding sporopollenin CYP enzymes, we found specific homologs of CYP704B1, which catalyzes the hydroxylation of long-chain fatty acids in all land plant clades. Homologs of this gene were not found in the sampled algae. We found homologs of CYP703A2 responsible for hydroxylation of mid-length-chain fatty acids only in the vascular plants and the mosses *Physcomitrella patens* and *Sphagnum fallax* but not in the liverwort *Marchantia polymorpha* or any algae (Figure 6B). This suggests that the use of lauric acid derivatives in sporopollenin biosynthesis originated in the bryophytes.

Despite the lack of specific CYP704B1 and CYP703A2 homologs, sporopollenin is found in algae. One possible explanation for this is that the oxidation reactions for sporopollenin biosynthesis are carried out by multifunctional CYP86 clan enzymes in algae, which when duplicated after the colonization of land became specialized for the biosynthesis of sporopollenin and other biopolymers.

Conserved Enzymes for Production of Secondary Monomers Found in Suberin and Sporopollenin Are Limited to Angiosperms and Land Plants, Respectively

The final stages of sporopollenin biosynthesis require polyketide synthases (PKSA/B) and tetraketide pyrone reductases (TKPR1/2) to produce tetraketide pyrones. No homologs of PKSA/B or TKPR1/2 were identified in the sampled algae (Figure 6B); our analysis suggests that tetraketide pyrones may be terrestrial plant-specific sporopollenin monomers.

Suberin and sporopollenin also contain various alcohols, the production of which requires fatty acid reductase (FAR) enzymes. In all but three of the angiosperm species (*Spirodela polyrhiza*, *Solanum tuberosum*, and *Malus domestica*), we found homologs of *Arabidopsis* FAR2, specific to sporopollenin biosynthesis. Homologs of *Arabidopsis* FARs specialized for suberin production (FAR1/4/5) were only found in the Brassicaceae (Figure 5). However, FAR family genes are found in all land plant taxa; those species missing FAR1/2/4/5 contain other FARs, which are more similar to *Arabidopsis* FARs of unknown or alternative function. This could mean that there has been significant sequence divergence from an ancestral specialized FAR. Or, that earlier-diverging land plants make use of multifunctional FARs for the production of these two polymers. Another possibility is that earlier-diverging land plant suberin and sporopollenin do not contain alcohols.

The use of very-long-chain fatty acids as suberin monomers means that a fatty acid elongation process is necessary. In angiosperms the enzymes responsible are 3-ketoacyl-CoA synthases (KCS2/20). As with FARs, we found homologs of *Arabidopsis* KCS genes in algae and land plants. However, in all species except dicotyledonous angiosperms and one monocot these genes are more related to alternative *Arabidopsis* KCS genes than KCS2/20 (Figure 5B).

After very-long-chain fatty acid production and oxidation, hydroxy fatty acids destined for cutin and suberin polymers are combined with glycerol-3-phosphate at the *sn*-2 position in a reaction catalyzed by glycerol 3-phosphate acyltransferase (GPAT) enzymes. GPAT4/6/8 gene products catalyze the transfer of long-chain fatty acids, while GPAT5/7 catalyze the transfer of suberin-specific, very-long-chain fatty acids. Despite the lack of specific genes encoding enzymes of the preceding oxidation and reduction steps, we found at least one homolog of *Arabidopsis* GPAT4/6/8 genes in all land plants (but none in algae) (Figure 5B). Consistent with evidence of reduced chain-length biopolymer monomers in earlier-diverging land plants [69] and presence of suberin only in vascular plants, specific homologs of GPAT5/7 were found only in the angiosperms and gymnosperms, and a single lycophyte (Figure 5B). It may be that the remaining vascular plants contain homologs of GPAT genes not yet characterized as suberin related.

The pattern of homology that we have described for structural molecule biosynthetic pathways supports the hypothesis that cutin appeared for the first time in the earliest land plants [71]. LACS genes already present in algae for the initial activation of free fatty acids were likely involved in cutin synthesis in these species. Indeed, homologous genes encoding LACS are present in almost all species in this study. Similarly, the final enzymes in the cutin biosynthesis pathway (GPAT4/6/8) are encoded by all land plant genomes analyzed (Figure 5B). However, non-seed plants, although containing genes belonging to the CYP86, CYP77, KCS, and FAR gene families, do not contain genes homologous to specific CYP86A/B and CYP77A subfamilies, or to KCS2/20 and FAR1/4/5.This suggests that oxidation reactions in cutin and suberin biosynthesis across taxa have been significantly modified since their origin in bryophytes and vascular plants, respectively, and the specific pattern of homolog presence of angiosperm biosynthetic genes in non-angiosperm plants suggest that the intermediate steps of the angiosperm-like pathway may have evolved in stages. The presence of chlorophyte algal genes in orthogroups containing *Arabidopsis* KCS and FAR genes suggest that, when suberin biosynthesis did develop, it was achieved by the co-option of pre-existing cutin biosynthetic CYP genes in combination with ancient elongase and reductase enzymes, the latter of which may have been multifunctional and also involved in sporopollenin biosynthesis. However, the lack of

Table 1. Results of Search for Homologs of Investigated Genes in Transcriptomes of 305 Non-angiosperm Plants and Algae from the One Thousand Plant Transcriptomes Initiative

Metabolic Pathway	Gene ID	*Arabidopsis* Accession	Charophytes	Liverworts	Mosses	Hornworts	Lycophytes	Ferns	Gymnosperms
Gibberellin biosynthesis	KO	AT5G25900	0/15	14/15[a]	15/15[a]	10/13[a]	10/15[a]	15/15[a]	15/15[a]
	KAO	AT1G05160; AT2G32440	0/19	9/16[a]	0/21	9/12[a]	9/18[a]	16/16[a]	16/17[a]
	GA20ox	AT5G07200; AT5G51810; AT1G60980	0/16	8/21[a]	19/22[a]	1/12[a]	7/16[a]	16/17[a]	16/16[a]
	GA3ox	AT1G15550; AT1G80340	0/16	0/20	7/17[a]	5/12[a]	6/18[a]	11/17[a]	6/23[a]
	GA2ox19	AT2G34555; AT1G30040	0/16	0/19	0/19	0/12	0/16	0/21	10/17[a]
	GA2ox20	AT4G21200; AT1G50960	0/15	5/19[a]	4/18[a]	4/12[a]	2/17[a]	1/21[a]	18/18[a]
	CYP714	AT5G24910; AT5G24900	0/18	0/17	0/16	0/12	0/15	0/17	0/16
	GAMT	AT4G26420; AT5G56300	8/8[a]	14/15[a]	18/18[a]	12/12[a]	16/16[a]	19/19[a]	15/15[a]
Brassinosteroid biosynthesis	DET2	AT2G38050	0/15	0/15	0/15	0/14	0/15	0/15	0/15
	CYP85A	AT5G38970; AT3g30180	0/22	0/16	0/16	0/12	0/16	0/20	0/19
	CYP90A1	AT5G05690	0/15	0/15	0/15	0/12	11/15[a]	15/15[a]	13/15[a]
	CYP90B1	AT3G50660	0/15	0/15	0/15	0/12	0/15	15/15[a]	15/15[a]
	CYP90C/D	AT4G36380; AT3G13730	0/20	0/18	0/22	0/12	0/15	0/22	24/24[a]
	CYP734A1	AT2G26710	0/15	0/15	0/14	0/12	0/15	0/15	0/15
Cutin/suberin biosynthesis	LACS	AT2G47240; AT1G64400; AT1G49430	22/22[a]	21/21[a]	24/24[a]	12/13[a]	18/18[a]	30/30[a]	32/33[a]
	CYP86A	AT5G58860; AT2G45970	0/17	5/18[a]	0/17	3/12[a]	1/17[a]	5/17[a]	2/26[a]
	HOTHEAD	AT1G72970	0/15	0/14	0/15	0/10	0/12	0/15	15/15[a]
	CYP77A	AT3G10570; AT5G04660	0/16	0/17	0/17	0/12	0/16	0/17	0/19
	GPAT4/6/8	AT1G01610; AT2G38110; AT4G00400	0/0	1/16[a]	0/19	0/10	18/18[a]	21/23[a]	29/30[a]
	KCS2/20	AT1G04220; AT5G43760	0/17	0/16	0/19	0/10	0/15	0/20	0/19
	FAR1/4/5	AT5G22500; AT3G44540; AT3G44550	0/16	0/10	0/18	0/10	0/15	0/20	0/18
	GPAT5/7	AT3G11430; AT5G06090	0/2	0/16	0/17	0/10	0/18	0/17	0/20
Sporopollenin	ACH	AT1G01710	15/15[a]	15/15[a]	15/15[a]	12/12[a]	15/15[a]	15/15[a]	15/15[a]
	CYP703A2	AT1G01280	0/15	1/15[a]	14/15[a]	2/12[a]	12/15[a]	15/15[a]	15/15[a]
	CYP704B1	AT1G69500	0/15	14/15[a]	15/15[a]	9/12[a]	11/15[a]	15/15[a]	2/15[a]
	ACOS5	AT1G62940	0/15	3/15[a]	3/15[a]	2/15[a]	5/15[a]	15/15[a]	2/15[a]
	FAR2	AT3G11980	0/15	0/10	0/15	0/10	0/15	0/15	0/15
	TKPR	AT4G35420; AT1G68540	0/17	7/20[a]	9/18[a]	4/11[a]	15/17[a]	23/23[a]	24/24[a]
	PKS	AT1G02050; AT4G34850	0/18	0/19	0/16	0/11	0/17	0/15	0/18

The numerator indicates the number of species in which transcripts encoding the enzyme (or enzyme group) were found. The denominator indicates the number of transcriptomes of each taxon analyzed. A maximum of 15 unique transcriptomes per taxon were analyzed, but note that the number of species hits will exceed 15 when more than one gene is responsible for a specified reaction, for example, in the case of LACS.

[a] Homologs of the gene/gene group in question were found in the corresponding transcriptomes.

Arabidopsis-like FAR1/4/5 genes in non-Brassicaceae shows that significant modifications to suberin biosynthesis have occurred even within the angiosperms. With the exception of specific FAR2 enzymes, all land plant clades contain homologs of the enzymes required for sporopollenin biosynthesis. The missing homologs in algae suggest that algal sporopollenin is produced using alternative enzymes and may differ in its composition.

Testing Metabolic Gains on a Broader Selection of Species Using 1KP Transcriptomes

To test our results from the orthogroup genome analysis, and to compensate for the relative lack of genome sequences available for non-angiosperm land plants and algae, we carried out an additional analysis of transcriptomes released by the One Thousand Plant Transcriptomes Initiative [34]. This also allowed the analysis to be extended to hornworts, for which there are no available genome sequences. While transcriptome analysis cannot be used to draw conclusions about gene absences, it can be used to confirm the presence of gene homologs. A total of 305 transcriptomes were analyzed, comprising 81 gymnosperms, 78 ferns, 20 lycophytes, 37 mosses, 29 liverworts, 14 hornworts, and 46 charophytes. For each of the metabolic pathways that we analyzed in detail in this paper, the presence of homologs of genes associated with metabolic reactions were tracked across the transcriptome dataset (Table 1). Comparison of the transcriptome results with the genome analysis (Figures 3, 4, 5, and 6) reveals broad corroboration for the conclusions drawn regarding the evolution of metabolic pathways for GA and brassinosteroid biosynthesis/inactivation and the biosynthesis of structural polymers, thereby strengthening the conclusions. The transcriptome data also revealed additional detail and new information, described in the subsequent sections.

For GA biosynthesis, the transcriptome data provide full support for the results drawn from the genome data with one minor exception: we did not find GA3ox transcripts in liverworts (Table 1), whereas this gene was identified in the genome of *Marchantia polymorpha* (Figure 3). The transcriptome data also reveal that hornworts are more similar to liverworts than to mosses with respect to GA biosynthesis: hornwort transcriptomes contain homologs of KAO, which are missing from moss genomes and transcriptomes (Table 1; Figure 3) [46, 47].

The transcriptome data also provide additional information about the evolution of GA inactivation enzymes: homologs of GAMT genes were identified in the transcriptomes of charophyte algae as well as land plants (Table 1). Additionally, although homologs of GA2-oxidase genes acting on C20 GAs were not found in the genomes of *M. polymorpha*, *S. fallax*, or *P. patens* (Figure 3), they were identified in the transcriptomes of five other liverwort species, four other moss species, and four hornworts (Table 1). The presence of an additional inactivation mechanism in bryophytes provides further support for the hypothesis that GAs are likely present in the bryophytes even though this has yet to be demonstrated experimentally. The presence of GAMT homologs in charophyte algae transcriptomes indicates that these genes are more ancient in origin than previously discussed.

For the brassinosteroid biosynthesis pathway, the transcriptome data were in broad agreement with the genome analysis (Figure 4; Table 1) with one substantive exception: we were unable to identify transcripts of DET2 in the transcriptome data despite the clear presence of DET2 gene homologs in all plant and algal genomes investigated (Figure 4). We also did not find transcripts of CYP85A or CYP734A1 genes in any species, although these are present in the genomes of gymnosperms (Figure 4). The transcriptome data support the absence of genes for brassinosteroid biosynthesis in the bryophytes, including hornworts. Additionally, the transcriptome data extend our understanding of the evolution of brassinosteroid biosynthetic genes in the CYP90 family. Genome analysis showed the presence of CYP90A1 and CYP90B1 in ferns and genes for the complete pathway are identified first in gymnosperms (Figure 4). Transcriptome analysis corroborated these findings, as well as that of lycophyte genes forming a sister group to the clade of genes containing gymnosperm and angiosperm CYP90C1 and CYP90D1 (Table 1). Additionally, the fern transcripts were similar to the lycophytes with respect to CYP90C1 and CYP90D1. It seems likely that these genes in ferns and lycophytes represent a more generic version (perhaps with differing substrate specificity) of specific CYP90C/D genes in gymnosperms and angiosperms. Finally, lycophyte transcriptomes were shown to contain homologs of CYP90A1, which were not identified in the genome analysis. Combining these results further supports the stepwise evolution hypothesized for the brassinosteroid biosynthesis pathway found in angiosperms, beginning with the evolution of CYP90A1 and genes similar to CYP90C/D in lycophytes, both of which are maintained in ferns with the addition of CYP90B1, before the completion of the biosynthetic pathway with the evolution of CYP90D1 in gymnosperms.

Finally, the transcriptome data are in full agreement with the genome analysis with respect to sporopollenin biosynthesis (Figure 6; Table 1) apart from the lack of PKS transcripts identified in any land plant taxon. The presence of transcripts involved in cutin and suberin biosynthesis were more scattered—we were unable to identify transcripts of several genes—specifically GPAT5/7 in vascular plants, HOTHEAD in ferns, and GPAT4/6/8 in bryophytes—which were identified in the genome analysis (Figure 5). Transcriptome analysis of cutin and suberin biosynthetic genes does, however, identify homologs of CYP86A genes in the liverworts, hornworts, lycophytes, and ferns, which suggests that the early stages of biosynthesis for these structural compounds are more similar across land plants than indicated by the genome analysis (Figure 5). Additionally, a single liverwort was found to contain a homolog of CYP703A2, which was not identified in the genome of *M. polymorpha* (Figure 5).

DISCUSSION

We have developed a scalable computational pipeline to systematically analyze and compare metabolic capabilities encoded in genomes and transcriptomes and have used it to advance our knowledge of the evolution of metabolic pathways in the Chloroplastida. The approach was validated by its ability to identify known metabolic innovations associated with plant evolution. The analysis allowed us to resolve several additional metabolic innovations in relation to land plant evolution.

Our analysis of enzymes involved in GA and brassinosteroid biosynthesis strongly supports the view that the biosynthetic capability for these two hormones is correlated with the transition of plants to land but that non-spermatophyte plants do not contain the full, or conserved, pathways for synthesis and inactivation of bioactive forms of the hormones as recognized in angiosperms (Figure 4; Table 1). Moreover, our analysis suggests that, while the complete angiosperm-like GA biosynthetic pathway is found in all vascular plants, the brassinosteroid biosynthetic pathway shows a stepwise increase in complexity with the evolution of CYP90A1 in lycophytes (Table 1) and CYP90B1 in ferns (Figure 4; Table 1) before the completion of the pathway in the gymnosperms (Figure 4; Table 1). The latter is a surprising observation because brassinosteroids have been found across plant taxa—in angiosperms, gymnosperms, ferns, lycophytes, the moss *P. patens*, the liverwort *M. polymorpha*, and the chlorophyte *C. vulgaris* [52, 72]. This may indicate that brassinosteroid production in algae, bryophytes, lycophytes, and ferns is carried out by enzymes that differ from those used in spermatophytes.

A surprising finding was the presence of genes in bryophytes encoding GA biosynthetic and inactivation enzymes (Figure 3; Table 1) despite the fact that recognized bioactive forms of GA have not been detected in the few bryophytes that have been analyzed [47, 50]. A plausible interpretation of this is that bioactive GAs are present at very low concentrations in bryophytes (as is the case with brassinosteroids [52]), and this has hindered previous attempts at their identification. Mass spectrometry instrumentation has increased substantially in sensitivity in recent years so this would be worth revisiting. A key feature for the regulatory function of phytohormones is the capacity to rapidly inactivate the bioactive forms of the molecules using specific enzymes. Overall, the angiosperms were found to have a greater capacity for various kinds of GA and brassinosteroid inactivation than non-angiosperms, and homologs of genes encoding the angiosperm brassinosteroid inactivation mechanisms were not found in non-seed plants (Figures 3 and 4; Table 1). Combined with the lack of some GA inactivation enzymes identified in non-seed plants, this indicates that the majority of known angiosperm inactivation mechanisms for both hormones are relatively recent innovations in land plant evolution. Non-seed plants must either utilize alternative inactivation processes or a different set of genes to carry out brassinosteroid inactivation reactions. As indicated by the presence of a GAMT gene in *Marchantia polymorpha* (Figure 3), and GA 2-oxidases acting on C20 GAs in bryophytes (Table 1), bryophytes may be limited to GA inactivation via these mechanisms. The lack of conserved inactivation mechanisms between taxa for both phytohormones indicates that different plant groups make use of alternative or divergent enzymes for homeostasis of GAs and brassinosteroids. The presence of GAMT homologs in charophyte algae transcriptomes indicates that these genes are more ancient in origin than previously discussed and suggests that GAMTs in land plants evolved from similar methyltransferases in algae, likely with alternative substrates, given the lack of homologs of GA biosynthetic genes in algae.

The analysis also revealed details about the likely evolution of several structural polymers. The capability for biosynthesis of precursor molecules required for the production of cutin/suberin monomers were present prior to the evolution of land plants, with the LACS gene encoding the enzyme long-chain acyl-CoA synthetase required for fatty acid activation present in the genomes of sampled land plants from all clades as well as in charophyte and chlorophyte algal genomes (Figure 5B). However, even though cutin is thought to be present in all vascular plants, the CYP enzymes involved in angiosperm cutin and suberin biosynthesis appear to have evolved in stages, with only CYP86A present in all land plant taxa (Table 1), followed by HOTHEAD in the ferns and gymnosperms and CYP77A in the angiosperms (Figure 5; Table1). For sporopollenin biosynthesis, homologs of the ACOS5 gene responsible for fatty acid activation were not found in algae but were present in all land plants, while the ACH enzymes, which remove CoA, were found in charophyte algae and land plants (Figure 6B). CYP enzymes involved in sporopollenin biosynthesis are conserved only in land plants (Figure 6; Table1), but sporopollenin has been identified in chlorophyte algae—suggesting that alternative CYP enzymes (likely members of the CYP86 clan) are responsible for sporopollenin biosynthesis in these species.

Taken together, the ubiquitous presence in land plants and algae of genes encoding enzymes acting early in the biosynthetic pathways for suberin, cutin, and sporopollenin suggests that none of these biopolymer biosynthetic pathways developed *de novo* in the land plants. It may be that certain enzymes involved in cutin and suberin biosynthesis evolved directly from those involved in sporopollenin biosynthesis; a common evolutionary origin of all biopolymers has previously been hypothesized [68, 69]. However, homologs of the angiosperm enzymes responsible for intermediate and end-point reactions are often absent from earlier-diverging land plant genomes (Figures 5B and 6B), even when these plants are known to contain such biopolymers. There are several possible explanations for this. It may be that different mechanisms, or at least alternative enzymes (for example differing in substrate specificity or activity), are utilized for biopolymer production in earlier-diverging land plants. The enzyme families in question tend to be expanded in the angiosperm species in this analysis; this diversity may be linked to subfunctionalization conferring increases in substrate specificity or catalytic efficiency in later-diverging land plants. It is plausible that these differences result in alternative monomer composition of biopolymers in divergent plant taxa, a phenomenon that has already been observed between some species [73, 74]. Another possibility is that the same reactions, leading to the same end products, are catalyzed by enzymes with common ancestry but that exhibit sequence divergence between earlier- and later-diverging land plants, and as such are not identified as homologous. Finally, the presence of similar biopolymers but absence of homologous genes across divergent plant taxa could be the result of convergent evolution, which is a common theme in land plant evolution [75, 76] and has already been identified in the biosynthesis of the biopolymer lignin, where lycophytes and spermatophytes have independently developed the ability to produce monomers derived from sinapyl alcohol [77].

Limitations of the Approach

The data presented here demonstrate the power of a systematic comparative analysis of large genome and transcriptome datasets to reveal new understanding about the evolution of biological pathways and processes. Nevertheless, there are some limitation to the approach. These stem mainly from the reliance on sequence homology as a tool for annotating genes with metabolic functions. First, in the case that two organisms do not have homologous genes for specific enzymes in common it is difficult to speculate as to how this reflects their biochemistry. For example, although cutin is biosynthesized in all land plants, angiosperm and non-angiosperm species do not share the specific cytochrome P450 enzymes used by higher plants (Figure 5). Without experimental analysis, it is impossible to determine whether this is indicative of a difference in gene function between these species (and non-angiosperm plants produce alternative cutin monomers) or whether the lack of homology is caused by superficial divergence at the sequence level (and enzyme function may in fact be conserved across land plants). Conversely, while sequence-based homology can identify candidates for functionally analogous genes across species, experimental analysis is also required to confirm these predictions. This may be particularly true for the cytochrome P450 genes, which exist in large, highly interrelated gene families that have been subject to rapid expansion and subfunctionalization over the course of plant evolution [7, 15].

Second, plants exhibit a huge range of specialized metabolism, and it is reasonable to expect that non-angiosperm plants, which are relatively understudied, have the capacity for metabolic functions that are thus far undiscovered. The use of a *de novo* metabolic annotation tool trained across plant and non-plant organisms allows for the detection of metabolic reactions in non-angiosperms where homologs may be present in microbes or fungi rather than angiosperms; however, this method does not have the power to detect previously unstudied metabolic pathways. Increasing knowledge of non-angiosperm plant metabolism will require extensive experimental and computational investigation. Metabolic profiling over different conditions and growth stages can be used in combination with flux analysis to elucidate the compounds produced by individual species and their role in the wider metabolic network. Building up such a knowledge base of metabolism in non-angiosperm plants is the only way to improve the accuracy of metabolic pathway inference and would have additional far-reaching benefits, for example, in improving the prospects for using non-angiosperm plants as experimentally tractable test beds for metabolic engineering strategies.

Finally, gene homology inferences are dependent on a set of high-quality genomes as their foundation. Plant genomes tend to be larger and more complex than animal genomes, and the high incidence of repetitive elements affects the quality of genome assembly and annotation [78]. Incorrect construction of gene models may lead to the inference of gaps in gene presence in species where there are none. This is particularly a problem for gymnosperm genomes, which can be up to 30 Gb and are rich in repetitive DNA, mostly transposable elements [79]. It may be that these issues with gymnosperm genome assembly are the cause of some missing homologs in individual genomes in this study—for example, in Figure 5 several gymnosperms are missing genes that are identified in all other species analyzed. We have attempted to mitigate this problem in this study by maximizing our sample size, making use of the majority of genome sequences available for non-angiosperm species and establishing further gene presences using transcriptome analysis of an additional 305 species. Furthermore, our findings are limited to identifying evolutionary trends across whole plant taxa, as opposed to basing conclusions on the presence or absence of genes in individual species.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Production of metabolic pathway databases
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Analysis of metabolic pathways
  - Identification of metabolic transitions
  - Transcriptome analysis
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2020.02.086.

ACKNOWLEDGMENTS

N.C. is grateful for a studentship from the Oxford Systems Biology Doctoral Training Program funded by the EPSRC. S.K. and D.M.E. were supported by the European Union's Horizon 2020 research and innovation program under grant agreement number 637765. S.K. is a Royal Society University Research Fellow. A.J.H. was funded by the George Grosvenor Freeman Fellowship by Examination in Sciences, Magdalen College (Oxford).

AUTHOR CONTRIBUTIONS

N.C. and L.J.S. conceived the project and developed the methodology; N.C. wrote code and carried out the investigation and formal analysis; L.J.S. provided supervision; S.K. and D.M.E. provided software and computing resources; J.M., L.D., and A.J.H. provided transcriptome resources; N.C. wrote the manuscript and prepared the figures; L.J.S., L.D., J.M., A.J.H., and D.M.E. reviewed and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 17, 2019
Revised: February 4, 2020
Accepted: February 27, 2020
Published: March 26, 2020

REFERENCES

1. de Vries, J., Stanton, A., Archibald, J.M., and Gould, S.B. (2016). Streptophyte terrestrialization in light of plastid evolution. Trends Plant Sci. *21*, 467–476.
2. Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z., Schneider, H., and Donoghue, P.C.J. (2018). The timescale of early land plant evolution. Proc. Natl. Acad. Sci. USA *115*, E2274–E2283.
3. Rensing, S.A. (2018). Great moments in evolution: the conquest of land by plants. Curr. Opin. Plant Biol. *42*, 49–54.
4. Berens, M.L., Berry, H.M., Mine, A., Argueso, C.T., and Tsuda, K. (2017). Evolution of hormone signaling networks in plant defense. Annu. Rev. Phytopathol. *55*, 401–425.
5. Davies, K.M., Albert, N.W., Zhou, Y., and Schwinn, K.E. (2018). Functions of flavonoid and betalain pigments in abiotic stress tolerance in plants. Ann. Plant Rev. Published online April 20, 2018. https://doi.org/10.1002/9781119312994.apr0604.
6. Niklas, K.J., Cobb, E.D., and Matas, A.J. (2017). The evolution of hydrophobic cell wall biopolymers: from algae to angiosperms. J. Exp. Bot. *68*, 5261–5269.
7. Nelson, D., and Werck-Reichhart, D. (2011). A P450-centric view of plant evolution. Plant J. *66*, 194–211.
8. Basler, G., Fernie, A.R., and Nikoloski, Z. (2018). Advances in metabolic flux analysis toward genome-scale profiling of higher organisms. Biosci. Rep. *38*. Published online November 23, 2018. https://doi.org/10.1042/BSR20170224.
9. Rensing, S.A. (2017). Why we need more non-seed plant models. New Phytol. *216*, 355–360.
10. Gomes de Oliveira Dal'Molin, C., and Nielsen, L.K. (2018). Plant genome-scale reconstruction: from single cell to multi-tissue modelling and omics analyses. Curr. Opin. Biotechnol. *49*, 42–48.
11. Moreira, T.B., Lima, J.M., Coca, G.C., and Williams, T.C.R. (2019). Insights into the spatial and temporal organisation of plant metabolism from network flux analysis. Theor. Exp. Plant Physiol. *31*, 215–226.
12. Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E.A., Kamisugi, Y., et al. (2008).

The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science *319*, 64–69.

13. Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., et al. (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. Nat. Commun. *5*, 3978.

14. Bowman, J.L., Kohchi, T., Yamato, K.T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., et al. (2017). Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. Cell *171*, 287–304.

15. Chae, L.; Kim, T.; Nilo-Poyanco, R.; Rhee, S.Y. Genomic signatures of specialized metabolism in plants. Science 344, 510–513.

16. Mueller, L.A., Zhang, P., and Rhee, S.Y. (2003). AraCyc: a biochemical pathway database for *Arabidopsis*. Plant Physiol. *132*, 453–460.

17. Nishiyama, T., Sakayama, H., de Vries, J., Buschmann, H., Saint-Marcoux, D., Ullrich, K.K., Haas, F.B., Vanderstraeten, L., Becker, D., Lang, D., et al. (2018). The *Chara* genome: secondary complexity and implications for plant terrestrialization. Cell *174*, 448–464.

18. Banks, J.A.; Nishiyama, T.; Hasebe, M.; Bowman, J.L.; Gribskov, M.; DePamphilis, C.; Albert, V.A.; Aono, N.; Aoyama, T.; Ambrose, B.A., et al. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science 332, 960–963.

19. Li, F.W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.M., Eily, A., Koppers, N., Kuo, L.Y., Li, Z., et al. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nat. Plants *4*, 460–472.

20. Van Ghelder, C., Parent, G.J., Rigault, P., Prunier, J., Giguère, I., Caron, S., Stival Sena, J., Deslauriers, A., Bousquet, J., Esmenjaud, D., and MacKay, J. (2019). The large repertoire of conifer NLR resistance genes includes drought responsive and highly diversified RNLs. Sci. Rep. *9*, 11614.

21. Caspi, R., Billington, R., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P.E., Ong, Q., Ong, W.K., et al. (2018). The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. *46* (D1), D633–D639.

22. Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A.K., Nilo-Poyanco, R., Bernard, T., et al. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. Plant Physiol. *173*, 2041–2059.

23. Karp, P.D., Latendresse, M., Paley, S.M., Krummenacker, M., Ong, Q.D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P., et al. (2016). Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Brief. Bioinform. *17*, 877–890.

24. Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., Shi, C., Wang, J., Liu, W., Liang, X., et al. (2016). Draft genome of the living fossil *Ginkgo biloba*. Gigascience *5*, 49.

25. Wan, T., Liu, Z.M., Li, L.F., Leitch, A.R., Leitch, I.J., Lohaus, R., Liu, Z.J., Xin, H.P., Gong, Y.B., Liu, Y., et al. (2018). A genome for gnetophytes and early evolution of seed plants. Nat. Plants *4*, 82–89.

26. Neale, D.B., McGuire, P.E., Wheeler, N.C., Stevens, K.A., Crepeau, M.W., Cardeno, C., Zimin, A.V., Puiu, D., Pertea, G.M., Sezen, U.U., et al. (2017). The Douglas-Fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. G3 (Bethesda) *7*, 3157–3167.

27. Stevens, K.A., Wegrzyn, J.L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Paul, R., Gonzalez-Ibeas, D., Koriabine, M., Holtz-Morris, A.E., et al. (2016). Sequence of the sugar pine megagenome. Genetics *204*, 1613–1626.

28. Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E., Liechty, J.D., et al. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. *15*, R59.

29. Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature *497*, 579–584.

30. Xu, Z., Xin, T., Bartels, D., Li, Y., Gu, W., Yao, H., Liu, S., Yu, H., Pu, X., Zhou, J., et al. (2018). Genome analysis of the ancient tracheophyte *Selaginella tamariscina* reveals evolutionary features relevant to the acquisition of desiccation tolerance. Mol. Plant *11*, 983–994.

31. Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. *40*, D1178–D1186.

32. Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T., Sun, W., Li, X., Xu, Y., Zhang, Y., et al. (2019). Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. Cell *179*, 1057–1067.

33. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

34. Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., et al.; One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature *574*, 679–685.

35. Van Der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

36. Rozentsvet, O.A. (2004). Comparative examination of distribution of phospholipids and a betaine lipid DGTS in tropical fern species. Biochem. Syst. Ecol. *32*, 303–311.

37. Rozentsvet, O.A., Dembitsky, V.M., and Saksonov, S.V. (2000). Occurrence of diacylglyceryltrimethylhomoserines and major phospholipids in some plants. Phytochemistry *54*, 401–407.

38. Riekhof, W.R., Naik, S., Bertrand, H., Benning, C., and Voelker, D.R. (2014). Phosphate starvation in fungi induces the replacement of phosphatidylcholine with the phosphorus-free betaine lipid diacylglyceryl-N,N,N-trimethylhomoserine. Eukaryot. Cell *13*, 749–757.

39. Novoselov, S.V., Rao, M., Onoshko, N.V., Zhi, H., Kryukov, G.V., Xiang, Y., Weeks, D.P., Hatfield, D.L., and Gladyshev, V.N. (2002). Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*. EMBO J. *21*, 3681–3693.

40. Rodman, J., Soltis, P., Soltis, D., Sytsma, K., and Karol, K. (1998). Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. Am. J. Bot. *85*, 997.

41. Clausen, M., Kannangara, R.M., Olsen, C.E., Blomstedt, C.K., Gleadow, R.M., Jørgensen, K., Bak, S., Motawie, M.S., and Møller, B.L. (2015). The bifurcation of the cyanogenic glucoside and glucosinolate biosynthetic pathways. Plant J. *84*, 558–573.

42. Fang, C., Fernie, A.R., and Luo, J. (2019). Exploring the diversity of plant metabolism. Trends Plant Sci. *24*, 83–98.

43. Halkier, B.A., and Gershenzon, J. (2006). Biology and biochemistry of glucosinolates. Annu. Rev. Plant Biol. *57*, 303–333.

44. Wang, C., Liu, Y., Li, S.S., and Han, G.Z. (2015). Insights into the origin and evolution of the plant hormone signaling machinery. Plant Physiol. *167*, 872–886.

45. Hayashi, K., Horie, K., Hiwatashi, Y., Kawaide, H., Yamaguchi, S., Hanada, A., Nakashima, T., Nakajima, M., Mander, L.N., Yamane, H., et al. (2010). Endogenous diterpenes derived from ent-kaurene, a common gibberellin precursor, regulate protonema differentiation of the moss *Physcomitrella patens*. Plant Physiol. *153*, 1085–1097.

46. Zi, J., Mafu, S., and Peters, R.J. (2014). To gibberellins and beyond! Surveying the evolution of (di)terpenoid metabolism. Annu. Rev. Plant Biol. *65*, 259–286.

47. Miyazaki, S., Hara, M., Ito, S., Tanaka, K., Asami, T., Hayashi, K.-I., Kawaide, H., and Nakajima, M. (2018). An ancestral gibberellin in a moss *Physcomitrella patens*. Mol. Plant *11*, 1097–1100.

48. Kumar, S., Kempinski, C., Zhuang, X., Norris, A., Mafu, S., Zi, J., Bell, S.A., Nybo, S.E., Kinison, S.E., Jiang, Z., et al. (2016). Molecular diversity of

terpene synthases in the liverwort *Marchantia polymorpha*. Plant Cell *28*, 2632–2650.

49. Noguchi, C., Miyazaki, S., Kawaide, H., Gotoh, O., Yoshida, Y., and Aoyama, Y. (2018). Characterization of moss ent-kaurene oxidase (CYP701B1) using a highly purified preparation. J. Biochem. *163*, 69–76.

50. Hirano, K., Nakajima, M., Asano, K., Nishiyama, T., Sakakibara, H., Kojima, M., Katoh, E., Xiang, H., Tanahashi, T., Hasebe, M., et al. (2007). The GID1-mediated gibberellin perception mechanism is conserved in the Lycophyte *Selaginella moellendorffii* but not in the Bryophyte *Physcomitrella patens*. Plant Cell *19*, 3058–3079.

51. Huang, Y., Wang, X., Ge, S., and Rao, G.Y. (2015). Divergence and adaptive evolution of the gibberellin oxidase genes in plants genome evolution and evolutionary systems biology. BMC Evol. Biol. *15*. Published online September 2015. https://doi.org/10.1186/s12862-015-0490-2.

52. Yokota, T., Ohnishi, T., Shibata, K., Asahina, M., Nomura, T., Fujita, T., Ishizaki, K., and Kohchi, T. (2017). Occurrence of brassinosteroids in non-flowering land plants, liverwort, moss, lycophyte and fern. Phytochemistry *136*, 46–55.

53. Cheon, J., Fujioka, S., Dilkes, B.P., and Choe, S. (2013). Brassinosteroids regulate plant growth through distinct signaling pathways in *Selaginella* and *Arabidopsis*. PLoS ONE *8*, e81938.

54. Ohnishi, T., Szatmari, A.M., Watanabe, B., Fujita, S., Bancos, S., Koncz, C., Lafos, M., Shibata, K., Yokota, T., Sakata, K., et al. (2006). C-23 hydroxylation by *Arabidopsis* CYP90C1 and CYP90D1 reveals a novel shortcut in brassinosteroid biosynthesis. Plant Cell *18*, 3275–3288.

55. Rieu, I., Eriksson, S., Powers, S.J., Gong, F., Griffiths, J., Woolley, L., Benlloch, R., Nilsson, O., Thomas, S.G., Hedden, P., and Phillips, A.L. (2008). Genetic analysis reveals that C19-GA 2-oxidation is a major gibberellin inactivation pathway in *Arabidopsis*. Plant Cell *20*, 2420–2436.

56. He, J., Chen, Q., Xin, P., Yuan, J., Ma, Y., Wang, X., Xu, M., Chu, J., Peters, R.J., and Wang, G. (2019). CYP72A enzymes catalyse 13-hydrolyzation of gibberellins. Nat. Plants *5*, 1057–1065.

57. Zhu, Y., Nomura, T., Xu, Y., Zhang, Y., Peng, Y., Mao, B., Hanada, A., Zhou, H., Wang, R., Li, P., et al. (2006). ELONGATED UPPERMOST INTERNODE encodes a cytochrome P450 monooxygenase that epoxidizes gibberellins in a novel deactivation reaction in rice. Plant Cell *18*, 442–456.

58. Zhang, Y., Zhang, B., Yan, D., Dong, W., Yang, W., Li, Q., Zeng, L., Wang, J., Wang, L., Hicks, L.M., and He, Z. (2011). Two *Arabidopsis* cytochrome P450 monooxygenases, CYP714A1 and CYP714A2, function redundantly in plant development through gibberellin deactivation. Plant J. *67*, 342–353.

59. Magome, H., Nomura, T., Hanada, A., Takeda-Kamiya, N., Ohnishi, T., Shinma, Y., Katsumata, T., Kawaide, H., Kamiya, Y., and Yamaguchi, S. (2013). CYP714B1 and CYP714B2 encode gibberellin 13-oxidases that reduce gibberellin activity in rice. Proc. Natl. Acad. Sci. USA *110*, 1947–1952.

60. Varbanova, M., Yamaguchi, S., Yang, Y., McKelvey, K., Hanada, A., Borochov, R., Yu, F., Jikumaru, Y., Ross, J., Cortes, D., et al. (2007). Methylation of gibberellins by *Arabidopsis* GAMT1 and GAMT2. Plant Cell *19*, 32–45.

61. Turk, E.M., Fujioka, S., Seto, H., Shimada, Y., Takatsuto, S., Yoshida, S., Denzel, M.A., Torres, Q.I., and Neff, M.M. (2003). CYP72B1 inactivates brassinosteroid hormones: an intersection between photomorphogenesis and plant steroid signal transduction. Plant Physiol. *133*, 1643–1653.

62. Poppenberger, B., Fujioka, S., Soeno, K., George, G.L., Vaistij, F.E., Hiranuma, S., Seto, H., Takatsuto, S., Adam, G., Yoshida, S., and Bowles, D. (2005). The UGT73C5 of *Arabidopsis thaliana* glucosylates brassinosteroids. Proc. Natl. Acad. Sci. USA *102*, 15253–15258.

63. Marsolais, F., Boyd, J., Paredes, Y., Schinas, A.-M., Garcia, M., Elzein, S., and Varin, L. (2007). Molecular and biochemical characterization of two brassinosteroid sulfotransferases from *Arabidopsis*, AtST4a (At2g14920) and AtST1 (At2g03760). Planta *225*, 1233–1244.

64. Delwiche, C.F.; Graham, L.E.; Thomson, N. Lignin-like compounds and sporopollenin in Coleochaete, an algal model for land plant ancestry. Science 245, 399–401.

65. Delwiche, C.F., and Cooper, E.D. (2015). The evolutionary origin of a terrestrial flora. Curr. Biol. *25*, R899–R910.

66. He, X., Dai, J., and Wu, Q. (2016). Identification of sporopollenin as the outer layer of cell wall in microalga *Chlorella protothecoides*. Front. Microbiol. *7*, 1047.

67. Vishwanath, S.J., Delude, C., Domergue, F., and Rowland, O. (2015). Suberin: biosynthesis, regulation, and polymer assembly of a protective extracellular barrier. Plant Cell Rep. *34*, 573–586.

68. Niklas, K.J., Cobb, E.D., and Matas, A.J. (2017). The evolution of hydrophobic cell wall biopolymers: from algae to angiosperms. J. Exp. Bot. *68*, 5261–5269.

69. Fich, E.A., Segerson, N.A., and Rose, J.K.C. (2016). The plant polyester cutin: biosynthesis, structure, and biological roles. Annu. Rev. Plant Biol. *67*, 207–233.

70. Kondo, S., Hori, K., Sasaki-Sekimoto, Y., Kobayashi, A., Kato, T., Yuno-Ohta, N., Nobusawa, T., Ohtaka, K., Shimojima, M., and Ohta, H. (2016). Primitive extracellular lipid components on the surface of the charophytic alga *Klebsormidium flaccidum* and their possible biosynthetic pathways as deduced from the genome sequence. Front. Plant Sci. *7*, 952.

71. Domínguez, E., Heredia-Guerrero, J.A., and Heredia, A. (2015). Plant cutin genesis: unanswered questions. Trends Plant Sci. *20*, 551–558.

72. Bajguz, A. (2009). Isolation and characterization of brassinosteroids from algal cultures of *Chlorella vulgaris Beijerinck* (Trebouxiophyceae). J. Plant Physiol. *166*, 1946–1949.

73. Caldicott, A.B., and Eglinton, G. (1976). Cutin acids from bryophytes: An ω-1 hydroxy alkanoic acid in two liverwort species. Phytochemistry *15*, 1139–1143.

74. Fernández, V., Guzmán-Delgado, P., Graça, J., Santos, S., and Gil, L. (2016). Cuticle structure in relation to chemical composition: Re-assessing the prevailing model. Front. Plant Sci. *7*, 427.

75. Huang, R., O'Donnell, A.J., Barboline, J.J., and Barkman, T.J. (2016). Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes. Proc. Natl. Acad. Sci. USA *113*, 10613–10618.

76. Sage, R.F., Christin, P.-A., and Edwards, E.J. (2011). The C(4) plant lineages of planet Earth. J. Exp. Bot. *62*, 3155–3169.

77. Weng, J.-K., Akiyama, T., Bonawitz, N.D., Li, X., Ralph, J., and Chapple, C. (2010). Convergent evolution of syringyl lignin biosynthesis via distinct pathways in the lycophyte *Selaginella* and flowering plants. Plant Cell *22*, 1033–1045.

78. Claros, M.G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. Biology (Basel) *1*, 439–459.

79. De La Torre, A.R., Birol, I., Bousquet, J., Ingvarsson, P.K., Jansson, S., Jones, S.J.M., Keeling, C.I., MacKay, J., Nilsson, O., Ritland, K., et al. (2014). Insights into conifer giga-genomes. Plant Physiol. *166*, 1724–1732.

80. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

81. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

82. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

83. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2017). GenBank. Nucleic Acids Res. *45* (D1), D37–D42.

84. Wegrzyn, J.L., Staton, M.A., Street, N.R., Main, D., Grau, E., Herndon, N., Buehler, S., Falk, T., Zaman, S., Ramnath, R., et al. (2019). Cyberinfrastructure to improve forest health and productivity: The role of

tree databases in connecting genomes, phenomes, and the environment. Front. Plant Sci. *10*, 813.

85. Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y.C., Sjödin, A., Van de Peer, Y., Jansson, S., Hvidsten, T.R., and Street, N.R. (2015). The Plant Genome Integrative Explorer Resource: New Phytol. *208*, 1149–1156 PlantGenIE.org.

86. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47* (D1), D506–D515.

87. Jeske, L., Placzek, S., Schomburg, I., Chang, A., and Schomburg, D. (2019). BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. *47* (D1), D542–D549.

88. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Chlamydomonas reinhardtii genome	Phytozome	N/A
Volvox carteri genome	Phytozome	N/A
Coccomyxa subellipsoidea C-169 genome	Phytozome	N/A
Micromonas pusilla CCMP1545 genome	Phytozome	N/A
Micromonas sp. RCC299 v3.0 genome	Phytozome	N/A
Chromochloris zofingiensis genome	Phytozome	N/A
Ostreococcus lucimarinus genome	Phytozome	N/A
Klebsormidium nitens genome	GenBank	GenBank: DF236950-DF238763
Chara braunii genome	[17]	https://bioinformatics.psb.ugent.be/gdb/Chara_braunii/
Mesotaenium endlicherianum genome	[32]	NCBI BioProject: PRJNA543678
Spirogloea muscicola genome	[32]	NCBI BioProject: PRJNA543679
Marchantia polymorpha genome	Phytozome	N/A
Physcomitrella patens genome	Phytozome	N/A
Sphagnum fallax genome	Phytozome	N/A
Selaginella moellendorffii genome	Phytozome	N/A
Selaginella tamariscina genome	[30]	NCBI Biosample: SAMN07840812
Isoetes echinospora transcriptome	GenBank	GenBank: GGKY00000000.1
Salvinia cucullata genome	Fernbase	https://www.fernbase.org/
Azolla filiculoides genome	Fernbase	https://www.fernbase.org/
Ginkgo biloba genome	[24]	https://doi.org/10.5524/100613
Gnetum montanum genome	[25]	https://doi.org/10.5061/dryad.0vm37
Pseudotsuga menziesii genome	TreeGenes	https://treegenesdb.org/
Pinus lambertiana genome	TreeGenes	https://treegenesdb.org/
Pinus taeda genome	ConGenIE	http://congenie.org/
Picea abies genome	ConGenIE	http://congenie.org/
Picea glauca transcriptome	GenBank	GenBank: GGJI00000000.1
Amborella trichopoda genome	Phytozome	N/A
Spirodela polyrhiza genome	Phytozome	N/A
Zostera marina genome	Phytozome	N/A
Asparagus officinalis genome	Phytozome	N/A
Ananas comosus genome	Phytozome	N/A
Musa acuminata genome	Phytozome	N/A
Brachypodium distachyon genome	Phytozome	N/A
Hordeum vulgare genome	Phytozome	N/A
Triticum aestivum genome	Phytozome	N/A
Oryza sativa genome	Phytozome	N/A
Oropetium thomaeum genome	Phytozome	N/A
Setaria italica genome	Phytozome	N/A
Sorghum bicolor genome	Phytozome	N/A
Zea mays genome	Phytozome	N/A
Chenopodium quinoa genome	Phytozome	N/A
Daucus carota genome	Phytozome	N/A
Lactuca sativa genome	Phytozome	N/A
Helianthus annuus genome	Phytozome	N/A

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Solanum lycopersicum genome	Phytozome	N/A
Solanum tuberosum genome	Phytozome	N/A
Mimulus guttatus genome	Phytozome	N/A
Olea europaea genome	Phytozome	N/A
Vitis vinifera genome	Phytozome	N/A
Eucalyptus grandis genome	Phytozome	N/A
Fragaria vesca genome	Phytozome	N/A
Malus domestica genome	Phytozome	N/A
Prunus persica genome	Phytozome	N/A
Cucumis sativus genome	Phytozome	N/A
Glycine max genome	Phytozome	N/A
Medicago truncatula genome	Phytozome	N/A
Cicer arietinum genome	Phytozome	N/A
Linum usitatissimum genome	Phytozome	N/A
Ricinus communis genome	Phytozome	N/A
Manihot esculenta genome	Phytozome	N/A
Citrus clementina genome	Phytozome	N/A
Citrus sinensis genome	Phytozome	N/A
Gossypium raimondii genome	Phytozome	N/A
Theobroma cacao genome	Phytozome	N/A
Carica papaya genome	Phytozome	N/A
Brassica oleracea capitata genome	Phytozome	N/A
Eutrema salsugineum genome	Phytozome	N/A
Arabidopsis lyrata genome	Phytozome	N/A
Arabidopsis thaliana genome	Phytozome	N/A
Boechera stricta genome	Phytozome	N/A
Capsella rubella genome	Phytozome	N/A
Capsella grandiflora genome	Phytozome	N/A
Klebsormidium nitens PGDB	This manuscript	N/A
Chara braunii PGDB	This manuscript	N/A
Marchantia polymorpha PGDB	This manuscript	N/A
Physcomitrella patens PGDB	This manuscript	N/A
Selaginella moellendorffii PGDB	This manuscript	N/A
Salvinia cucullata PGDB	This manuscript	N/A
Picea glauca PGDB	This manuscript	N/A
Arabidopsis thaliana PGDB	[16]	https://www.plantcyc.org/downloads
Encephalartos barteri	1KP Initiative	GNQG
Stangeria eriopus	1KP Initiative	KAWQ
Dioon edule	1KP Initiative	WLIC
Cycas micholitzii	1KP Initiative	XZUY
Ginkgo biloba	1KP Initiative	SGTW
Gnetum montanum	1KP Initiative	GTHK
Ephedra sinica	1KP Initiative	VDAO
Agathis macrophylla	1KP Initiative	ACWS
Chamaecyparis lawsoniana	1KP Initiative	AIGO
Pseudolarix amabilis	1KP Initiative	AQFM
Nothotsuga longibracteata	1KP Initiative	AREG
Widdringtonia cedarbergensis	1KP Initiative	AUDE
Picea engelmannii	1KP Initiative	AWQB

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Lepidothamnus sp.	1KP Initiative	BBDD
Austrotaxus spicata	1KP Initiative	BTTS
Platycladus orientalis	1KP Initiative	BUWV
Manoao colensoi	1KP Initiative	CDFR
Tetraclinis sp.	1KP Initiative	CGDN
Cryptomeria japonica	1KP Initiative	DSXO
Pinus radiata	1KP Initiative	DZQM
Torreya taxifolia	1KP Initiative	EFMS
Prumnopitys andina	1KP Initiative	EGLZ
Cunninghamia lanceolata	1KP Initiative	ESYX
Pilgerodendron uviferum	1KP Initiative	ETCJ
Taxodium distichum	1KP Initiative	FHST
Dacrycarpus compactus	1KP Initiative	FMWZ
Calocedrus decurrens	1KP Initiative	FRPM
Tsuga heterophylla	1KP Initiative	GAMH
Cedrus libani	1KP Initiative	GGEA
Cephalotaxus harringtonia	1KP Initiative	GJTI
Diselma archeri	1KP Initiative	GKCZ
Sequoia sempervirens	1KP Initiative	HBGV
Acmopyle pancheri	1KP Initiative	HILW
Cryptomeria japonica	1KP Initiative	HOUF
Torreya nucifera	1KP Initiative	HQOM
Amentotaxus argotaenia	1KP Initiative	IAJW
Callitris gracilis	1KP Initiative	IFLI
Pinus parviflora	1KP Initiative	IIOL
Pseudotsuga wilsoniana	1KP Initiative	IOVS
Dacrydium balansae	1KP Initiative	IZGN
Pinus ponderosa	1KP Initiative	JBND
Neocallitropsis pancheri	1KP Initiative	JDQB
Phyllocladus hypophyllus	1KP Initiative	JRNA
Keteleeria evelyniana	1KP Initiative	JUWL
Parasitaxus usta	1KP Initiative	JZVE
Sundacarpus amarus	1KP Initiative	KLGF
Pinus jeffreyi	1KP Initiative	MFTM
Microcachrys tetragona	1KP Initiative	MHGD
Agathis robusta	1KP Initiative	MIXZ
Thujopsis dolabrata	1KP Initiative	NKIN
Cathaya argyrophylla	1KP Initiative	NPRL
Metasequoia glyptostroboides	1KP Initiative	NRXL
Cunninghamia lanceolata	1KP Initiative	OUOI
Papuacedrus papuana	1KP Initiative	OVIJ
Halocarpus bidwillii	1KP Initiative	OWFC
Glyptostrobus pensilis	1KP Initiative	OXGJ
Saxegothaea conspicua	1KP Initiative	QCGM
Sequoiadendron giganteum Glaucum	1KP Initiative	QFAE
Falcatifolium taxoides	1KP Initiative	QHBI
Cupressus dupreziana	1KP Initiative	QNGJ
Taiwania cryptomerioides	1KP Initiative	QSNJ
Callitris macleayana	1KP Initiative	RMMV

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Falcatifolium taxoides	1KP Initiative	ROWR
Wollemia nobilis	1KP Initiative	RSCE
Podocarpus coriaceus	1KP Initiative	SCEB
Fokienia hodginsii	1KP Initiative	UEVI
Nageia nagi	1KP Initiative	UUJS
Thuja plicata	1KP Initiative	VFYZ
Retrophyllum minus	1KP Initiative	VGSX
Abies lasiocarpa	1KP Initiative	VSRH
Larix speciosa	1KP Initiative	WVWN
Taxus baccata	1KP Initiative	WWSS
Cephalotaxus harringtonia	1KP Initiative	WYAJ
Athrotaxis cupressoides	1KP Initiative	XIRK
Podocarpus rubens	1KP Initiative	XLGK
Juniperus scopulorum	1KP Initiative	XMGP
Microbiota decussata	1KP Initiative	XQSG
Araucaria rulei	1KP Initiative	XTZO
Sciadopitys verticillata	1KP Initiative	YFZK
Pseudotaxus chienii	1KP Initiative	YLPM
Austrocedrus chilensis	1KP Initiative	YYPE
Tmesipteris parva	1KP Initiative	ALVQ
Botrypus virginianus	1KP Initiative	BEGM
Equisetum diffusum	1KP Initiative	CAPN
Danaea nodosa	1KP Initiative	DFHO
Sceptridium dissectum	1KP Initiative	EEAQ
Equisetum hyemale	1KP Initiative	JVSZ
Psilotum nudum	1KP Initiative	QGVS
Psilotum nudum	1KP Initiative	QVMR
Marattia attenuata	1KP Initiative	UGNK
Marattia sp.	1KP Initiative	UXCS
Ophioglossum petiolatum	1KP Initiative	WTJG
Osmundastrum cinnamomeum	1KP Initiative	BIVQ
Aspenium sp.	1KP Initiative	BMIF
Adiantum raddianum	1KP Initiative	BMJR
Polypodium glycyrrhiza	1KP Initiative	CJNT
Anemia tomentosa	1KP Initiative	CQPW
Azolla cf. caroliniana	1KP Initiative	CVEG
Gaga arizonica	1KP Initiative	DCDT
Thyrsopteris elegans	1KP Initiative	EWXK
Deparia lobato-crenata	1KP Initiative	FCHS
Pteris ensiformis	1KP Initiative	FLTD
Polystichum acrostichoides	1KP Initiative	FQGQ
Cyathea (Alsophila) spinulosa	1KP Initiative	GANB
Myriopteris rufa	1KP Initiative	GSXD
Polypodium hesperium	1KP Initiative	GYFU
Gymnocarpium dryopteris	1KP Initiative	HEGQ
Cystopteris utahensis	1KP Initiative	HNDZ
Onoclea sensibilis	1KP Initiative	HTFH
Polypodium hesperium	1KP Initiative	IXLH
Bolbitis repanda	1KP Initiative	JBLI

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Pilularia globulifera	1KP Initiative	KIIX
Asplenium platyneuron	1KP Initiative	KJZG
Cystopteris fragilis	1KP Initiative	LHLE
Dipteris conjugata	1KP Initiative	MEKP
Thelypteris acuminata	1KP Initiative	MROH
Dennstaedtia davallioides	1KP Initiative	MTGC
Vittaria appalachiana	1KP Initiative	NDUV
Lindsaea leariz	1KP Initiative	NOKI
Nephrolepis exaltata	1KP Initiative	NWWI
Homalosorus pycnocarpos	1KP Initiative	OCZL
Davallia fejeensis	1KP Initiative	OQWW
Phymatosorus grossus	1KP Initiative	ORJE
Lygodium japonicum	1KP Initiative	PBUU
Ceratopteris thalictroides	1KP Initiative	PIVW
Culcita macrocarpa	1KP Initiative	PNZO
Pteris vittata	1KP Initiative	POPJ
Asplenium nidus	1KP Initiative	PSKY
Hymenophyllum bivalve	1KP Initiative	QIAD
Osmundastrum cinnamomeum	1KP Initiative	RFMZ
Didymochlaena truncatula	1KP Initiative	RFRB
Ricciocarpos natans	1KP Initiative	RICC
Vittaria lineata	1KP Initiative	SKYV
Hymenophyllum cupressiforme	1KP Initiative	TRPJ
Crepidomanes venosum	1KP Initiative	TWFZ
Diplazium wichurae	1KP Initiative	UFJN
Pityrogramma trifoliata	1KP Initiative	UJTT
Pleopeltis polypodioides	1KP Initiative	UJWU
Osmunda sp.	1KP Initiative	UOMY
Athyrium filix-femina	1KP Initiative	URCP
Plagiogyria japonica	1KP Initiative	UWOD
Osmunda javanica	1KP Initiative	VIBO
Blechnum spicant	1KP Initiative	VITX
Lonchitis hirsuta	1KP Initiative	VVRN
Adiantum aleuticum	1KP Initiative	WCLG
Leucostegia immersa	1KP Initiative	WGTU
Cryptogramma acrostichoides	1KP Initiative	WQML
Argyrochosma nivea	1KP Initiative	XDDT
Sticherus lobatus	1KP Initiative	XDVM
Cystopteris fragilis	1KP Initiative	XXHP
Notholaena montieliae	1KP Initiative	YCKE
Lindsaea microphylla	1KP Initiative	YIXP
Woodsia scopulina	1KP Initiative	YJJY
Osmunda regalis	1KP Initiative	YKSS
Polypodium amorphum	1KP Initiative	YLJA
Cystopteris protrusa	1KP Initiative	YOWV
Woodsia ilvensis	1KP Initiative	YQEC
Phlebodium pseudoaureum	1KP Initiative	ZQYU
Parahemionitis cordata	1KP Initiative	ZXJO
Selaginella lepidophylla	1KP Initiative	ABIJ

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Huperzia myrsinites	1KP Initiative	CBAE
Lycopodium annotinum	1KP Initiative	ENQF
Huperzia squarrosa	1KP Initiative	GAON
Huperzia lucidula	1KP Initiative	GKAG
Huperzia selago	1KP Initiative	GTUO
Selaginella wallacei	1KP Initiative	JKAA
Selaginella willdenowii	1KP Initiative	KJYC
Selaginella selaginoides	1KP Initiative	KUXM
Selaginella apoda	1KP Initiative	LGDQ
Huperzia selago	1KP Initiative	NYBX
Isoetes tegetiformans	1KP Initiative	PKOX
Lycopodium deuterodensum	1KP Initiative	PQTO
Isoetes sp.	1KP Initiative	PYHZ
Lycopodiella appressa	1KP Initiative	ULKT
Pseudolycopodiella caroliniana	1KP Initiative	UPMJ
Diphasiastrum digitatum	1KP Initiative	WAFT
Dendrolycopodium obscurum	1KP Initiative	XNXF
Selaginella kraussiana	1KP Initiative	ZFGK
Selaginella stauntoniana	1KP Initiative	ZZOL
Blasia sp.	1KP Initiative	AEXY
Radula lindenbergiana	1KP Initiative	BNCU
Frullania spp.	1KP Initiative	CHJJ
Treubia lacunosa	1KP Initiative	FITN
Sphaerocarpos texanus	1KP Initiative	HERT
Marchantia paleacea	1KP Initiative	HMHL
Ptilidium pulcherrimum	1KP Initiative	HPXA
Marchantia paleacea	1KP Initiative	IHWO
Conocephalum conicum	1KP Initiative	ILBQ
Scapania nemorosa	1KP Initiative	IRBN
Barbilophozia barbata	1KP Initiative	JHFI
Marchantia polymorpha	1KP Initiative	JPYU
Porella navicularis	1KP Initiative	KRUQ
Schistochila sp.	1KP Initiative	LGOW
Metzgeria crassipilis	1KP Initiative	NRWZ
mixed species of Plagiochilaceae/Fissidentaceae	1KP Initiative	NWQC
Barbilophozia barbata	1KP Initiative	OFTV
Pellia sp. (cf. epiphylla (L.) Corda)	1KP Initiative	PIUF
Calypogeia fissa	1KP Initiative	RTMU
Monoclea gottschei	1KP Initiative	TFDQ
Marchantia emarginata	1KP Initiative	TFYI
Frullania sp.	1KP Initiative	TGKW
Lunularia cruciata	1KP Initiative	TXVB
Porella pinnata	1KP Initiative	UUHD
Ricciocarpos natans	1KP Initiative	WJLO
Bazzania trilobata	1KP Initiative	WZYK
Odontoschisma prostratum	1KP Initiative	YBQN
Pallavicinia lyellii	1KP Initiative	YFGP
Noteroclada confluens	1KP Initiative	YPSN

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Racomitrium elongatum	1KP Initiative	ABCD
Diphyscium foliosum	1KP Initiative	AWOI
Scouleria aquatica	1KP Initiative	BPSG
Orthotrichum lyellii	1KP Initiative	CMEQ
Thuidium delicatulum	1KP Initiative	EEMJ
Ceratodon purpureus	1KP Initiative	FFPD
Sphagnum lescurii	1KP Initiative	GOWD
Syntrichia princeps	1KP Initiative	GRKU
Buxbaumia aphylla	1KP Initiative	HRWG
Tetraphis pellucida	1KP Initiative	HVBQ
Leucodon julaceus	1KP Initiative	IGUH
Rhynchostegium serrulatum	1KP Initiative	JADL
Encalypta streptocarpa	1KP Initiative	KEFD
Stereodon subimponens	1KP Initiative	LNSF
Climacium dendroides	1KP Initiative	MIRS
Dicranum scoparium	1KP Initiative	NGTD
Pseudotaxiphyllum elegans	1KP Initiative	QKQO
Anomodon attenuatus	1KP Initiative	QMWB
Sphagnum palustre	1KP Initiative	RCBT
Racomitrium varium	1KP Initiative	RDOO
Leucobryum glaucum	1KP Initiative	RGKI
Takakia lepidozioides	1KP Initiative	SKQD
Polytrichum commune	1KP Initiative	SZYG
Calliergon cordifolium	1KP Initiative	TAVP
Neckera douglasii	1KP Initiative	TMAJ
Sphagnum recurvum	1KP Initiative	UHLI
Claopodium rostratum	1KP Initiative	VBMM
Leucobryum albidum	1KP Initiative	VMXJ
Aulacomnium heterostichum	1KP Initiative	WNGH
Loeskeobryum brevirostre	1KP Initiative	WSPM
Rosulabryum cf. capillare	1KP Initiative	XWHK
Physcomitrium sp.	1KP Initiative	YEPO
Hedwigia ciliata	1KP Initiative	YWNF
Leucodon brachypus	1KP Initiative	ZACW
Timmia austriaca	1KP Initiative	ZQRI
Atrichum angustatum	1KP Initiative	ZTHV
Phaeomegaceros coriaceus	1KP Initiative	AKXB
Leiosporoceros dussii	1KP Initiative	ANON
Anthoceros agrestis	1KP Initiative	BSNI
Nothoceros aenigmaticus	1KP Initiative	DXOU
Paraphymatoceros hallii	1KP Initiative	FAJB
Leiosporoceros dussii	1KP Initiative	FANS
Anthoceros angustus	1KP Initiative	IQJU
Phaeoceros carolinianus	1KP Initiative	RXRQ
Nothoceros vincentianus	1KP Initiative	TCBC
Anthoceros agrestis	1KP Initiative	TWUW
Megaceros flagellaris	1KP Initiative	UCRN
Phaeoceros carolinianus	1KP Initiative	WCZB
Phaeoceros carolinianus	1KP Initiative	WEEQ

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Phaeoceros carolinianus	1KP Initiative	ZFRE
Penium margaritaceum	1KP Initiative	AEKF
Entransia fimbria ta	1KP Initiative	BFIK
Cosmarium tinctum	1KP Initiative	BHBK
Desmidium aptogonum	1KP Initiative	DFDS
Closterium lunula	1KP Initiative	DRFX
Chaetosphaeridium globosum	1KP Initiative	DRGY
Netrium digitus	1KP Initiative	FFGR
Klebsormidium subtile	1KP Initiative	FQLP
Xanthidium antilopaeum	1KP Initiative	GBGT
Onychonema laeve	1KP Initiative	GGWH
Euastrum affine	1KP Initiative	GYRP
Spirogyra sp.	1KP Initiative	HAOX
Cosmarium broomei	1KP Initiative	HIDG
Cosmarium ochthodes	1KP Initiative	HJVM
Staurastrum sebaldi	1KP Initiative	ISHC
Cylindrocystis cushleckae	1KP Initiative	JOJQ
Gonatozygon kinahanii	1KP Initiative	KEYW
Nucleotaenium eifelense	1KP Initiative	KMNX
Micrasterias fimbriata	1KP Initiative	MCHJ
Zygnemopsis sp.	1KP Initiative	MFZO
Cosmarium granatum	1KP Initiative	MNNM
Pleurotaenium trabecula	1KP Initiative	MOYY
Chara vulgaris	1KP Initiative	MWXT
Mesotaenium kramstae	1KP Initiative	NBYP
Spirotaenia minuta	1KP Initiative	NNHQ
Coleochaete irregularis	1KP Initiative	QPDY
Bambusina borreri	1KP Initiative	QWFV
Cylindrocystis brebissonii	1KP Initiative	RPGL
Phymatodocis nordstedtiana	1KP Initiative	RPQV
Staurodesmus omearii	1KP Initiative	RPRU
Cosmocladium cf. constrictum	1KP Initiative	RQFE
Planotaenium ohtanii	1KP Initiative	SNOX
Cosmarium ochthodes	1KP Initiative	STKJ
Spirotaenia sp.	1KP Initiative	TPHT
Cylindrocystis sp.	1KP Initiative	VAZE
Coleochaete scutata	1KP Initiative	VQBJ
Staurodesmus convergens	1KP Initiative	WCQU
Mesotaenium endlicherianum	1KP Initiative	WDCW
Cosmarium subtumidum	1KP Initiative	WDGV
Zygnema sp.	1KP Initiative	WGMD
Mesotaenium braunii	1KP Initiative	WSJO
Roya obtusa	1KP Initiative	XRTZ
Cylindrocystis brebissonii	1KP Initiative	YOXI
Penium exiguum	1KP Initiative	YSQT
Mougeotia sp.	1KP Initiative	ZRMT
Software and Algorithms		
E2P2	[22]	https://gitlab.com/rhee-lab/E2P2/tree/master
Pathway Tools	[23]	RRID:SCR_013786; http://bioinformatics.ai.sri.com/ptools/

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
OrthoFinder	[33]	RRID:SCR_017118; https://github.com/davidemms/OrthoFinder
MAFFT	[80]	RRID:SCR_011811; https://mafft.cbrc.jp/alignment/software/
trimAl	[81]	RRID:SCR_017334; http://trimal.cgenomics.org/
IQ-TREE	[82]	RRID:SCR_017254; http://www.iqtree.org/
Metabolic Innovation Identifier Script	This manuscript	N/A

LEAD CONTACT AND MATERIALS AVAILABILITY

This study did not generate new unique reagents.

A key resources table is included. Datasets and code generated during this study are available in Figures S1, S2, S3, S4, S5, and S6; Table 1; Tables S1 and S2; and Data S1, S2, S3, S4, S5, and S6.

Requests for further information and resources should be directed to and will be fulfilled by the Lead Contact, Lee Sweetlove (lee.sweetlove@plants.ox.ac.uk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study used a total of 377 genome and transcriptome datasets. All angiosperm genome sequences used in the orthogroup analysis were downloaded from the Phytozome database (https://phytozome.jgi.doe.gov/). Non-angiosperm genome and transcriptome sequences used in the orthogroup analysis were either downloaded from Phytozome, GenBank [83], Fernbase [19], TreeGenes [84], or ConGenie [85], or from individual publications; specific sources for each species can be found in the key resources table. All 305 transcriptomes used in the transcriptome analysis were taken from the One Thousand Plant Transcriptomes Initiative [34], and unique identifiers for each species are documented in the key resources table.

METHOD DETAILS

Production of metabolic pathway databases

Metabolic pathway/genome databases (PGDBs) for seven streptophyte species (*Klebsormidium nitens*, *Chara braunii*, *Marchantia polymorpha*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Salvinia cucullata* and *Picea glauca*) were generated (Data S1) using the Pathway Tools software [23], which uses genome annotations to infer the presence of metabolic reactions in an organism and a rule-based approach to import entire metabolic pathways once certain thresholds of reaction evidence are reached (Data S5). The software was set up to utilize both MetaCyc and PlantCyc as reference metabolic pathway databases. Pathway Tools computes a likelihood score, incorporating information about the proportion of metabolic reactions in the pathway with corresponding annotations in an organism's genome, the uniqueness of these reactions across metabolic pathways and whether any are considered 'key' reactions to the pathway. A pathway is imported into the PGDB if the likelihood score is greater than a specified threshold value. This was set relatively low (0.15) to minimize the number of false negative pathway inferences. To be included in a PGDB, a pathway must also meet several further criteria: it must be a natural metabolic pathway (not genetically engineered), must not be missing any specified key reactions, and must have an expected taxonomic range matching the organism in question. The latter criterion is problematic for studies such as this which include earlier-diverging land plants and algae on which relatively little metabolic research has been carried out, increasing the likelihood of false negatives and limiting our ability to extend current knowledge beyond well-studied higher plants. Despite this taxonomy restriction, the Pathway Tools software also intrinsically reduces the number of false negative inferences caused by missing genome annotations – given a sufficiently high pathway likelihood score, complete metabolic pathways are incorporated into a PGDB, including any reactions for which an associated genome annotation has not been found. The low pathway likelihood score inclusion threshold maximizes this effect.

Efforts were made to further mitigate the possibility of false negative inferences by improving the genome annotations on which the PGDBs were based (Data S5). Prior to PGDB construction, the raw protein sequences of all chosen species were passed through the E2P2 software (Ensemble-Enzyme Prediction Pipeline [22]), a machine learning-based algorithm developed for the extraction of metabolism-specific information from genome sequences. The algorithm is trained on a set of more than 142,000 protein sequences, compiled from enzyme and non-enzyme protein sequences from any organism taken from SwissProt [86], BRENDA [87], MetaCyc or PlantCyc, with the stipulation that the entry must either have been manually curated or has experimental support [22]. E2P2 automatically annotates input sequences for each organism with EC numbers and MetaCyc-specific reaction identifiers (E2P2 annotations for each of the seven species are available in Data S1). Following genome annotation with E2P2, Pathway Tools was used to produce two PGDBs, individually based on the previously published genome annotation for each organism and the E2P2 annotation produced here. The union of the metabolic pathways inferred in each case became the final PGDB for each organism. It is important to note that since this method, like the majority of gene annotation efforts, is based on sequence homology and protein domain identification, it has little power to identify novel metabolic enzymes (and therefore reactions and pathways). Rather, this approach attempts to produce the most complete set of metabolic pathways possible for each organism, given the totality of existing knowledge of metabolic reactions.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of metabolic pathways

The PGDBs for each of the seven streptophyte species detailed above and the well-curated *Arabidopsis* PGDB [16] were consolidated into a binary presence/absence table for all metabolic pathways present in any species PGDB. To compare the patterns of metabolic pathway presence/absence across species, k-mediods clustering followed by an unsupervised clustering technique known as t-distributed stochastic neighbor embedding (t-SNE [35]) was carried out. For the clustering of species, t-SNE was run 500,000 times and the solution with the minimal final cost function value was chosen.

Identification of metabolic transitions

For a broader multi-species comparison, 64 plant and algal species in addition to the eight species described above were compared, taking the gene-reaction associations identified in the eight original species as the starting point. The additional species include seven chlorophyte algae, two further charophytes (*Spirogloea muscicola* and *Mesotaenium endlicherianum*), one further bryophyte (*Sphagnum fallax)*, two further lycophytes (*Selaginella tamariscina* and *Isoetes echinospora*), one further fern (*Azolla filiculoides*), six further gymnosperms and 45 further angiosperms. All angiosperm genome sequence data were taken from the Phytozome online repository [31]; sources for additional non-angiosperm species are identified in the key resources table. Genome sequence data was used for all additional species except the lycophyte *I. echinospora,* for which the transcriptome was used (available on NCBI's GenBank [83], GenBank: GGKY00000000.1). Genes for all 72 species were sorted into orthogroups using OrthoFinder software [33], which infers orthogroups of genes across species from sequence data (where an orthogroup is a collection of orthologous and paralogous genes descending from one gene in the common ancestor of all species included in the analysis). Orthogroups are produced using an all-versus-all DIAMOND [88] search followed by the normalization of the resulting scores for gene length and phylogenetic distance. A gene network is then constructed, weighted with the normalized scores and this network is clustered into orthogroups using Markov clustering (MCL). Orthogroups for all 72 species and corresponding orthogroup gene trees are available as Data S2 and Data S3. A script to query this orthogroup data based on *Arabidopsis* accession IDs is provided in Data S4.

Both the gene-reaction association data and the orthogroup data were processed using a Python script (available as Data S6) to identify pathways representing possible metabolic innovations or losses in land plant evolution. For every metabolic pathway across all representative species PGDBs, support for each metabolic reaction in the pathway in each of the genomes of the eight representative species was tabulated from the gene-reaction association data used by Pathway Tools. Equivalent support for each metabolic reaction in the 72 species on which orthogroups had been calculated was identified mainly on the basis of *Arabidopsis* annotations since *Arabidopsis* is the best studied species included in the analysis and likely has the highest quality genome annotation. If an *Arabidopsis* gene is annotated with a reaction, genes in other organisms found in the orthogroup corresponding to this *Arabidopsis* gene are considered functionally homologous and associated with the same reaction. If no *Arabidopsis* gene is annotated with a reaction, but at least two of the seven further representative species' genome annotations contain annotations for this reaction, then as long as an orthogroup exists which contains at least one such annotated gene from all the representative species, this orthogroup is considered associated with the reaction and corresponding genes in other organisms in this orthogroup are again considered functionally homologous.

Evidence for each metabolic pathway from both the gene-reaction associations and OrthoFinder were consolidated into a set of tables containing this information across the eight representative species (Data S5). Reaction presence was evaluated sequentially from earlier-diverging to later-diverging species. If this progression along the phylogeny was associated with a single change in the proportion of reactions evidenced (e.g., at a single point the proportion of reactions for which there is genetic evidence increases and remains high for the rest of the progression, or decreases and remains low), the metabolic pathway was flagged as involved in a possible metabolic innovation or loss in land plant evolution and sent to a results file containing all such pathways. The orthogroups associated with each reaction in each flagged metabolic pathway were then inspected to confirm that the evidence pattern held across the 64 further species included in the orthogroups. Evidence for each flagged metabolic pathway in each of the eight species annotated was scrutinised to remove false positive hits (e.g., where existing literature showed the identified pattern to be incorrect) and those for which limited reaction information was available, and to ensure that assumptions were not being made based on unreasonably large orthogroups. Some orthogroups are too large to assume that the spectrum of included genes may have similar function; in these situations gene trees for each orthogroup were investigated and genes were only considered candidates for functional homology if closer relationships were observed within subtrees. For the purposes of removing false positives, the gene trees were investigated when genes required for pathways fell in one of the largest 200 orthogroups (containing more than 860 genes across the 72 species). For the genes presented in the results section of this work, all gene trees of any size were inspected.

Transcriptome analysis

Seven blast databases were created from all available transcripts for charophytes, liverworts, mosses, hornworts, lycophytes, ferns and gymnosperms individually using DIAMOND. For each gene, we identified whether reverse blast hits of the *Arabidopsis* gene in question are present in the relevant species' transcriptomes. Genes from two species are considered to be reverse blast hits when each is the top-scoring DIAMOND result in a sequence comparison against the other species' genome/transcriptome. For the initial blast, the total number of DIAMOND results was set to 100, and the single top scoring genes from the top 15 scoring species were extracted for reverse-comparison in a second DIAMOND search back against the *Arabidopsis* genome. Transcripts from each set of species which pass this reciprocal test were added to the multiple sequence alignment of all the genes in the corresponding orthogroup using MAFFT [80] and the resulting alignments trimmed using trimAl [81] with the 'gappyout' option. We then built a phylogenetic tree from the new multiple sequence alignment using IQ-TREE [82] with the model 'LG+I+G4' and inspected each phylogeny manually to ensure that the transcripts identified are indeed most closely related to the original *Arabidopsis* gene in question (as opposed to forming an outgroup, or being homologs of alternative *Arabidopsis* genes).

DATA AND CODE AVAILABILITY

The published article contains all datasets, code and supporting information generated during this study.