11-11-2014

Phylotranscriptomic analysis of the origin and early diversification of land plants

Norman J. Wickett
Chicago Botanic Garden

Siavash Mirarab
The University of Texas at Austin

Nam Nguyen
The University of Texas at Austin

Tandy Warnow
The University of Texas at Austin

Eric Carpenter
University of Alberta

See next page for additional authors

Recommended Citation

Wickett, N., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M., Burleigh, J., Gitzendanner, M., Ruhfel, B., Wafula, E., Der, J., Graham, S., Mathews, S., Melkonian, M., Soltis, D., Soltis, P., Miles, N., Rothfels, C., Pokorny, L., Shaw, A., De Gironimo, L., Stevenson, D., Surek, B., Villarreal, J., Roure, B., Philippe, H., De Pamphilis, C., Chen, T., Deyholos, M., Baucom, R., & Kutchan, T. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America, 111* (45), E4859-E4868. https://doi.org/10.1073/pnas.1323926111

Phylotranscriptomic analysis of the origin and early diversification of land plants

Norman J. Wickett[a,b,1,2], Siavash Mirarab[c,1], Nam Nguyen[c], Tandy Warnow[c], Eric Carpenter[d], Naim Matasci[e,f], Saravanaraj Ayyampalayam[g], Michael S. Barker[f], J. Gordon Burleigh[h], Matthew A. Gitzendanner[h,i], Brad R. Ruhfel[h,j,k], Eric Wafula[l], Joshua P. Der[l], Sean W. Graham[m], Sarah Mathews[n], Michael Melkonian[o], Douglas E. Soltis[h,i,k], Pamela S. Soltis[h,i,k], Nicholas W. Miles[k], Carl J. Rothfels[p,q], Lisa Pokorny[p,r], A. Jonathan Shaw[p], Lisa DeGironimo[s], Dennis W. Stevenson[s], Barbara Surek[o], Juan Carlos Villarreal[t], Béatrice Roure[u], Hervé Philippe[u,v], Claude W. dePamphilis[l], Tao Chen[w], Michael K. Deyholos[d], Regina S. Baucom[x], Toni M. Kutchan[y], Megan M. Augustin[y], Jun Wang[z], Yong Zhang[v], Zhijian Tian[z], Zhixiang Yan[z], Xiaolei Wu[z], Xiao Sun[z], Gane Ka-Shu Wong[d,z,aa,2], and James Leebens-Mack[g,2]

[a]Chicago Botanic Garden, Glencoe, IL 60022; [b]Program in Biological Sciences, Northwestern University, Evanston, IL 60208; [c]Department of Computer Science, University of Texas, Austin, TX 78712; [d]Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada T6G 2E9; [e]iPlant Collaborative, Tucson, AZ 85721; [f]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721; [g]Department of Plant Biology, University of Georgia, Athens, GA 30602; [h]Department of Biology and [i]Genetics Institute, University of Florida, Gainesville, FL 32611; [j]Department of Biological Sciences, Eastern Kentucky University, Richmond, KY 40475; [k]Florida Museum of Natural History, Gainesville, FL 32611; [l]Department of Biology, Pennsylvania State University, University Park, PA 16803; [m]Department of Botany and [q]Department of Zoology, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; [n]Arnold Arboretum of Harvard University, Cambridge, MA 02138; [o]Botanical Institute, Universität zu Köln, Cologne D-50674, Germany; [p]Department of Biology, Duke University, Durham, NC 27708; [r]Department of Biodiversity and Conservation, Real Jardín Botánico-Consejo Superior de Investigaciones Cientificas, 28014 Madrid, Spain; [s]New York Botanical Garden, Bronx, NY 10458; [t]Department fur Biologie, Systematische Botanik und Mykologie, Ludwig-Maximilians-Universitat, 80638 Munich, Germany; [u]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Succursale Centre-Ville, Montreal, QC, Canada H3C 3J7; [v]CNRS, Station d' Ecologie Expérimentale du CNRS, Moulis, 09200, France; [w]Shenzhen Fairy Lake Botanical Garden, The Chinese Academy of Sciences, Shenzhen, Guangdong 518004, China; [x]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109; [y]Donald Danforth Plant Science Center, St. Louis, MO 63132; [z]BGI-Shenzhen, Bei shan Industrial Zone, Yantian District, Shenzhen 518083, China; and [aa]Department of Medicine, University of Alberta, Edmonton, AB, Canada T6G 2E1

Edited by Paul O. Lewis, University of Connecticut, Storrs, CT, and accepted by the Editorial Board September 29, 2014 (received for review December 23, 2013)

Reconstructing the origin and evolution of land plants and their algal relatives is a fundamental problem in plant phylogenetics, and is essential for understanding how critical adaptations arose, including the embryo, vascular tissue, seeds, and flowers. Despite advances in molecular systematics, some hypotheses of relationships remain weakly resolved. Inferring deep phylogenies with bouts of rapid diversification can be problematic; however, genome-scale data should significantly increase the number of informative characters for analyses. Recent phylogenomic reconstructions focused on the major divergences of plants have resulted in promising but inconsistent results. One limitation is sparse taxon sampling, likely resulting from the difficulty and cost of data generation. To address this limitation, transcriptome data for 92 streptophyte taxa were generated and analyzed along with 11 published plant genome sequences. Phylogenetic reconstructions were conducted using up to 852 nuclear genes and 1,701,170 aligned sites. Sixty-nine analyses were performed to test the robustness of phylogenetic inferences to permutations of the data matrix or to phylogenetic method, including supermatrix, supertree, and coalescent-based approaches, maximum-likelihood and Bayesian methods, partitioned and unpartitioned analyses, and amino acid versus DNA alignments. Among other results, we find robust support for a sister-group relationship between land plants and one group of streptophyte green algae, the Zygnematophyceae. Strong and robust support for a clade comprising liverworts and mosses is inconsistent with a widely accepted view of early land plant evolution, and suggests that phylogenetic hypotheses used to understand the evolution of fundamental plant traits should be reevaluated.

land plants | Streptophyta | phylogeny | phylogenomics | transcriptome

The origin of embryophytes (land plants) in the Ordovician period roughly 480 Mya (1–4) marks one of the most important events in the evolution of life on Earth. The early evolution of embryophytes in terrestrial environments was facilitated by numerous innovations, including parental protection for the developing embryo, sperm and egg production in multicellular protective structures, and an alternation of phases (often referred to as generations) in which a diploid sporophytic life history stage gives rise to a multicellular haploid gametophytic phase. With

Significance

Early branching events in the diversification of land plants and closely related algal lineages remain fundamental and unresolved questions in plant evolutionary biology. Accurate reconstructions of these relationships are critical for testing hypotheses of character evolution: for example, the origins of the embryo, vascular tissue, seeds, and flowers. We investigated relationships among streptophyte algae and land plants using the largest set of nuclear genes that has been applied to this problem to date. Hypothesized relationships were rigorously tested through a series of analyses to assess systematic errors in phylogenetic inference caused by sampling artifacts and model misspecification. Results support some generally accepted phylogenetic hypotheses, while rejecting others. This work provides a new framework for studies of land plant evolution.

Author contributions: N.J.W., S. Mirarab, T.W., S.W.G., M.M., D.E.S., P.S.S., D.W.S., M.K.D., J.W., G.K.-S.W., and J.L.-M. designed research; N.J.W., S. Mirarab, N.N., T.W., E.C., N.M., S.A., M.S.B., J.G.B., M.A.G., B.R.R., E.W., J.P.D., S.W.G., S. Mathews, M.M., D.E.S., P.S.S., N.W.M., C.J.R., L.P., A.J.S., L.D., D.W.S., B.S., J.C.V., B.R., H.P., C.W.d., T.C., M.K.D., M.M.A., J.W., Y.Z., Z.T., Z.Y., X.W., X.S., G.K.-S.W., and J.L.-M. performed research; S. Mirarab, N.N., T.W., N.M., S.A., M.S.B., J.G.B., M.A.G., E.W., J.P.D., S.W.G., S. Mathews, M.M., D.E.S., P.S.S., N.W.M., C.J.R., L.P., A.J.S., L.D., D.W.S., B.S., J.C.V., H.P., C.W.d., T.C., M.K.D., R.S.B., T.M.K., M.M.A., J.W., Y.Z., G.K.-S.W., and J.L.-M. contributed new reagents/analytic tools; N.J.W., S. Mirarab, N.N., E.C., N.M., S.A., M.S.B., J.G.B., M.A.G., B.R.R., E.W., B.R., H.P., and J.L.-M. analyzed data; N.J.W., S. Mirarab, T.W., S.W.G., M.M., D.E.S., D.W.S., H.P., G.K.-S.W., and J.L.-M. wrote the paper; and N.M. archived data.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. P.O.L. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the iplant datastore database, mirrors.iplantcollaborative.org/onekp_pilot, and the National Center for Biotechnology Information Sequence Read Archive, www.ncbi.nlm.nih.gov/sra [accession no. PRJEB4921 (ERP004258)].

[1]N.J.W. and S. Mirarab contributed equally to this work.

[2]To whom correspondence may be addressed. Email: nwickett@chicagobotanic.org, gane@ualberta.ca, or jleebensmack@plantbio.uga.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1323926111/-/DCSupplemental.

these and subsequent innovations, embryophytes diversified and the lineage ultimately came to dominate and significantly alter terrestrial environments (1–4). The origin of embryophytes was a pivotal event in evolutionary history that spawned the tremendous diversity of morphological, physiological, reproductive, and ecological traits we see in both the extant and fossil terrestrial flora. Moreover, colonization of land by plants greatly changed the global carbon cycle, drawing down atmospheric CO_2 concentrations (5) and forming the foundation of the vast majority of terrestrial ecosystems.

Subsequent innovations in embryophyte evolution greatly expanded the diversity of the terrestrial flora. The origin of vascular tissue and antidesiccation features in tracheophytes (vascular plants) established a more efficient system for the transport and retention of water, photosynthate, and other nutrients, as well as providing the cellular foundation for wood. Physiological innovations were accompanied by a shift in life history, from gametophytic to sporophytic dominance. The origin of the seed in the seed plant lineage greatly increased parental provisioning for the embryo, and the origin of the flower in the angiosperm lineage prompted a series of rapid radiations, yielding the most diverse group of extant plants.

Much of our current understanding of plant phylogeny has come from the study of plastid data (e.g., refs. 6–11), mitochondrial genes (e.g., refs. 12 and 13) and ribosomal gene analyses (e.g., refs. 14 and 15). The more recent application of phylogenomic analyses to large numbers of nuclear genes has generally supported previous hypotheses, but taxon sampling has been sparse and some inferred relationships remain controversial. Two fundamental questions persist with respect to the origin and diversification of embryophytes: (*i*) which streptophytic green algal lineage is most closely related to embryophytes, and (*ii*) what is the branching order among major embryophyte lineages? We aim to build on previous phylogenomic investigations of the earliest branching events in streptophyte evolution through increased sampling of taxa representing key lineages and innovations. Refined understanding of these events will inform investigations of traits that have contributed to key innovations in plant evolution.

Although the monophyly of Streptophyta (streptophytic green algae plus embryophytes) is well established (16–25), the inferred branching order of streptophytic algal lineages relative to embryophytes remains uncertain (26–30). Conflict among previous studies may derive from differing taxon and gene sampling and different methods of analysis. Within streptophytes, embryophytes, Charales, and Coleochaetales share derived, complex characteristics, including oogamous sexual reproduction, parental retention of the egg, apical growth with branching, and the presence of plasmodesmata in the gametophytic phase: pores in the cell wall allowing cytoplasmic transport of molecules between neighboring cells. Furthermore, the phragmoplast, a collection of microtubules and actin microfilaments that directs formation of the cell plates during cytokinesis, is shared among embryophytes, Charales, Coleochaetales, and at least some members of the Zygnematophyceae (31–33). A four-gene phylogeny that included markers from all three genomic compartments was consistent with the previously hypothesized sister-group relationship of Charales and embryophytes that, together, were sister to Coleochaetales (34). However, recent phylogenomic analyses based on complete plastome sequences, discrete plastome regions, ribosomal protein genes, and other nuclear genes have instead inferred that either Coleochaetales (35), Zygnematophyceae (8, 27–30, 36), or a clade including both lineages (28, 37, 38) are sister to embryophytes. These results imply that either complex characters, such as branching, parental retention of the egg, and plasmodesmata originated independently in the Charales, Coleochaetales, and embryophytes, or they originated once in a common ancestor and were subsequently lost in most lineages within the Zygnematophyceae.

Early events in the diversification of embryophytes gave rise to mosses, liverworts, and hornworts (collectively bryophytes) (25, 39–44). Virtually every possible hypothesis of branching order involving these groups has been proposed and supported by various data. Resolving this uncertainty has implications for understanding evolution of the heteromorphic alternation of life history phases shared by all embryophytes. Whereas all bryophyte lineages share a life history in which the haploid phase (gametophyte) is dominant, with a diploid phase (sporophyte) that is dependent on the maternal gametophyte, vascular plants instead have a dominant sporophytic phase. A grade of bryophytes would support the hypothesis that the gametophyte-dominant life cycle is plesiomorphic in embryophytes (45). In contrast, if bryophytes are monophyletic, it is equally likely that the common ancestor of all land plants was characterized by either a gametophyte-dominant or a sporophyte-dominant life cycle. Furthermore, fossil taxa with isomorphic life history phases (i.e., neither the sporophyte nor the gametophyte is dominant) have been described from the Rhynie Chert (46); our interpretation of the origin and evolution of plants with heteromorphic or isomorphic generations may be shaped by the resolution of bryophyte lineages. Indeed, bryophytes have been resolved as monophyletic in some analyses (41, 47), but analyses indicating a grade of liverworts, mosses, hornworts, with the latter as the sister group of tracheophytes (e.g., ref. 45), have been largely favored. This latter branching order has been supported by molecular phylogenetic analyses (25, 38, 43) and mitochondrial intron gains (42, 48), but it has also been rejected by several analyses. For example, mosses and liverworts have been resolved as monophyletic in phylogenetic analyses of complete plastomes (23, 49), multigene datasets (40), and morphological analyses (44). The position of the hornworts relative to a mosses+liverwort clade and tracheophytes, however, has varied in these studies, and sparse taxon sampling may have influenced resulting topologies.

Key tracheophyte relationships have also been revisited with genomic data, including investigations of relationships within and among lycophytes and monilophytes (49), the position of Gnetales within a monophyletic gymnosperm clade (50), and the branching order among angiosperm lineages. Increasingly, analyses of nuclear genes assembled from publicly available genome or transcriptome databases are being used to assess previously recalcitrant relationships within the green tree of life (27, 29, 35, 51–53).

Here, we present an analysis of 852 protein-coding nuclear genes for 103 taxa obtained by mining 92 streptophyte (algae and embryophytes) transcriptomes generated de novo, at least in part, for this study, plus 11 publicly available plant nuclear genome sequences. Whereas taxon sampling in phylogenomic analyses is generally sparse, the transcriptome data presented here greatly expand coverage across the green plant clade and sampling density within many key clades. We analyze these data using a comprehensive set of data-filtering and analytical approaches to assess whether inferred relationships are robust across analyses or possibly artifacts of data limitations or misspecification of evolutionary models used in phylogenetic inference algorithms.

Results and Discussion

Transcriptome Sequencing and Sorting. Protein sequences from 25 publicly available genomes were clustered into 27,054 operationally defined gene families using orthoMCL (54). Hidden Markov models (HMMs) were computed for each of these inferred gene-family circumscriptions or "orthogroups" (55) and used to assign transcript assemblies for 92 species (Table S1) to the appropriate orthogroups. To maintain focus on Streptophyta while avoiding oversampling some flowering plant clades, only 11 of the 25

publicly available sequenced genomes used to define orthogroups were included in our phylogenomic analyses (Table S1).

After filtering, multiple sequence alignments (MSAs) and gene trees were estimated for 9,610 gene families that included at least four taxa (transcriptome assemblies, unfiltered gene family alignments, and trees are available through the iPlant Data Store and can be accessed via iPlant Discovery Environment or at mirrors.iplantcollaborative.org/onekp_pilot). Of these, we identified 852 gene families that included at most one gene copy from at least 24 of the 25 sequenced genomes. These putatively single-copy gene families were used to estimate relationships (56) among the species included in Table S1. For those taxa where more than one sequence mapped to the same typically single-copy orthogroup, a consensus sequence was generated and retained if nucleotide divergence between the overlapping sequences was 5% or less; if divergence was greater than 5%, that species was not included in the MSA for that gene family. As a consequence, all filtered orthogroup MSAs included at most one sequence per taxon; a sequence for a particular taxon may have been missing from a single-copy gene family alignment because of lack of expression, gene loss, or putative lineage-specific duplication (Fig. S1 and Table S1).

Matrices and Analyses. Simultaneous alignment and tree estimation (SATé) (57) alignments of the 852 orthogroups were used to estimate phylogenetic relationships through supermatrix, supertree, and coalescent-based species tree approaches. The concatenated, untrimmed nucleotide supermatrix included 1,701,170 aligned sites and 50,715,288 nongap characters. Individual orthogroup matrices and the supermatrix were also filtered more stringently to investigate how missing data, highly divergent sequences (possible contaminants), and data type (nucleotides vs. inferred amino acids) influenced inferred relationships estimated using contrasting methods of analysis [RAxML (58) and PhyloBayes (59) supermatrix analyses, SuperFine (60) supertree analyses, and ASTRAL (61), a method designed to take into account gene tree incongruence resulting from incomplete lineage sorting between speciation events]. In total, we ran 69 analyses (Table S2) and compared results to assess robustness to variation in data-filtering and analysis strategies (see for example, Fig. 4).

All species-tree estimates were assessed for resolution of hypothesized relationships among focal clades, e.g., the identity of the sister group to embryophytes (land plants); relationships among bryophytes [Marchantiophyta (liverworts), Bryophyta (mosses), Anthocerotophyta (hornworts)]; and placement of Gnetales (Fig. 1). The tree estimates produced from most analyses were highly concordant and largely consistent with the relationships reflected in the maximum-likelihood (ML) tree estimated from nucleotides at the first and second codon positions (Figs. 2 and 3). However, differences among analytical approaches were observed with respect to the resolution of relationships that have been long-debated in the plant systematics literature (see below) (Fig. 4).

Some of the discordance among trees (i.e., strongly supported relationships that are incongruent among trees), derived from different methods of analysis, could be attributed to model misspecification. The most extreme contrast in inferred relationships was observed between analyses of nucleotide alignments including all three codon positions and analyses of only first and second nucleotide positions or those based on amino acid alignments (Fig. 4). The large variation observed in GC content at the third codon position (Figs. S2 and S3) is not accounted for in the ML analyses of nucleotide alignments under the GTR+Gamma substitution model. Therefore, in the following discussion we focus on results from analyses of first and second codon position and amino acid alignments.

Relationships Among Streptophytic Algal Lineages and Land Plants. In all analyses, Streptophyta are monophyletic, with a clade including Mesostigmatales, Chlorokybales, and *Spirotaenia* resolved as sister to all remaining streptophytes. The phylogenetic position of *Spirotaenia minuta* (sister to *Chlorokybus*) does not come as a surprise because previous analyses of *rbcL* and SSUrDNA datasets including three other species of *Spirotaenia* (including the type species, *Spirotaenia condensata*) showed that this genus does not belong in the Zygnematophyceae, but rather is affiliated with *Chlorokybus* and *Mesostigma* (62). Thus, taxonomic circumscription of *Spirotaenia* and traditional placement of all *Spirotaenia* species in the Zygnematophyceae are erroneously based on homoplasious morphological characters, including the shape of the chloroplast and sexual reproduction by conjugation. No analysis provided strong support for a sister relationship between Coleochaetales and embryophytes, and most analyses rejected a sister relationship between Charales and embryophytes (Fig. 4 and Fig. S4). Analyses of nucleotide data that included third positions offered weak support for Charales sister to embryophytes, but as mentioned above, this is likely an artifact of among-lineage variation in nucleotide frequencies at the third codon position (Fig. S2).

The results presented here provide strong support for a sister group relationship between Zygnematophyceae and embryophytes in analyses of amino acids and first and second codon positions (Figs. 2–4), a relationship that has been inferred in recent analyses of plastomes (8, 36) and a smaller set of nuclear gene sequences (27, 29). Whereas most individual gene trees did not provide strong support for any of the hypotheses illustrated

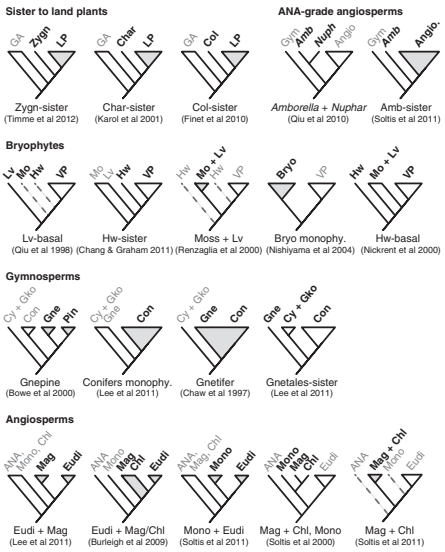

Fig. 1. Hypothesized land plant relationships evaluated for all matrices included in this study. Support for these hypotheses was evaluated across all 69 analyses performed in this study. For each hypothesis shown here, a "bar" (e.g., "GA" in the "Sister to land plants" block) indicates unspecified resolution within a grade. Dashed lines indicate lineages or grades with placements that are not relevant to the specified hypothesis. Note that differences with the nonfocal parts of the topology may exist in the given references (see below), and that additional studies may have recovered similar topologies. Abbreviations used: ANA, Amborellales, Nymphaeales, Austrobaileyales grade; Bryo, bryophytes; Char, Charales; Chl, Chloranthales; Col, Coleochaetales; Con, conifers; Cy, cycads; Eudi, eudicots; GA, green algae; Gko, *Ginkgo*; Gne, Gnetales; Gym, gymnosperms; Hw, hornworts; LP, land plants; Lv, liverworts; Mag, magnoliids; Mo, mosses; Mono, monocots; Pin, Pinaceae; VP, vascular plants; Zygn, Zygnematophyceae. Sister to land plants: Timme et al. (27), Karol et al. (34), Finet et al. (35), Bryophytes: Qiu et al. (42), Chang and Graham (38), Renzaglia et al. (44), Nishiyama et al. (41), Nickrent et al. (40); Gymnosperms: Bowe et al. (92), Lee et al. (52), Chaw et al. (91); Angiosperms: Qiu et al. (12), Burleigh et al. (15), Soltis et al. (14), Soltis et al. (105).

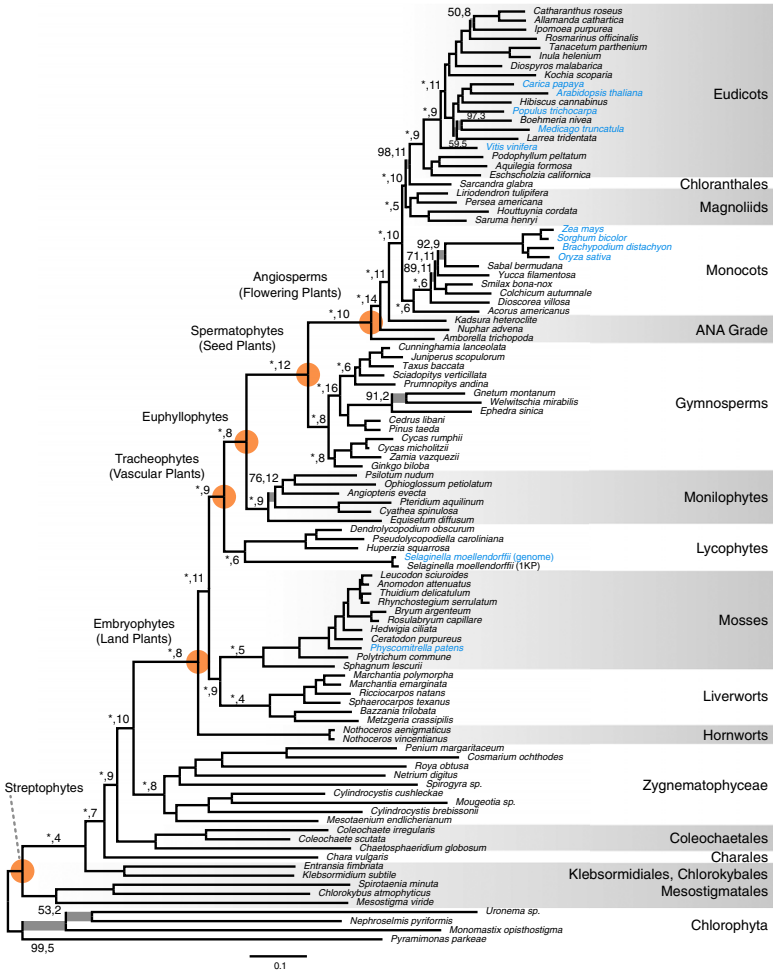

Fig. 2. ML phylogram inferred from concatenated alignments of first and second codon positions for 674 genes after gene alignments missing more than 50% of taxa were removed and sites gapped in more than 50% of taxa were filtered. Bootstrap values for the concatenated supermatrix analysis are shown on branches (* for 100%) along with the percentages of gene trees exhibiting significant conflict (bootstrap support >75%) with nodes discussed in the text. See Figs. S6 and S7 for bootstrap values and percentages of conflicting gene trees for all nodes in analyses of first and second codon positions and amino acids, respectively. Most nodes exhibited 100% bootstrap support in the concatenated analysis, but gray branches highlight nodes with less than 100% support. Gene models from sequenced genomes are indicated by blue text.

in Fig. 1, a small proportion of gene trees did exhibit well-supported conflict with each hypothesis (Figs. 2 and 3, and Fig. S3). This discordance was not unexpected and may be because of incomplete sorting of ancestral variation between speciation events represented by short internodes in the species phylogeny (63, 64) (Fig. 2). ASTRAL analyses (61) of gene trees estimated from amino acid alignments recovered strong support for Zygnematophyceae as sister to land plants (Fig. 4). ASTRAL analyses of in-frame nucleotide data, when first and second positions alone are considered, recovered the same relationship but with weaker support (Figs. 3 and 4); after filtering fragmentary sequences to improve gene tree resolution, we again recovered this relationship with high support (Fig. 4). As seen in our supermatrix and supertree analyses, ASTRAL analyses of nucleotide data including all codon positions recovered trees with weak support for *Chara* as the sister lineage to land plants. Again, this result is interpreted as an artifact of among-lineage variation in character-state frequencies.

Zygnematophyceae are a group of unicellular or filamentous streptophyte algae that sexually reproduce by conjugation, rather than flagellate cells (65). The absence of motile cells and plasmodesmata in Zygnematales may be interpreted as secondary reduction of morphological complexity following divergence from a common ancestor shared with Charales and Coleochaetales, which is consistent with their mode of reproduction (29). Phragmoplast presence and structure is also consistent with this interpretation of secondary loss, as they seem to be absent from most Zygnematophyceae, but simplified phragmoplasts have been characterized for the filamentous *Spirogyra* (31, 66), *Mougeotia* (33), and likely *Zygnema* (67). Fowke and Pickett-Heaps (31) suggested that the rudimentary phragmoplast seen in *Spirogyra* may represent an ancestral form, but placement of Zygnematophyceae as sister to land plants implies that a simplified (rather than ancestral) phragmoplast existed in the zygnematophycean stem lineage and was independently lost within the two major zygnematalean clades (Figs. 2 and 3). The possibility of independent origins of phragmoplasts in multiple streptophyte lineages appears unlikely; however, the phycoplast, a collection of microtubules serving a similar function in cytokinesis relative to the phragmoplast but forming parallel to the division plane (in contrast to the phragmoplast), did evolve independently in the lineage leading to the core chlorophytes (68, 69). Reports on the occurrence of phragmoplast-mediated cytokinesis in the ulvophycean chlorophytes *Trentepohlia* and *Cephaleuros* (70), however, should be interpreted with caution, as functional studies are lacking and structurally this system is more reminiscent of a rudimentary telophase spindle than a genuine streptophyte phragmoplast.

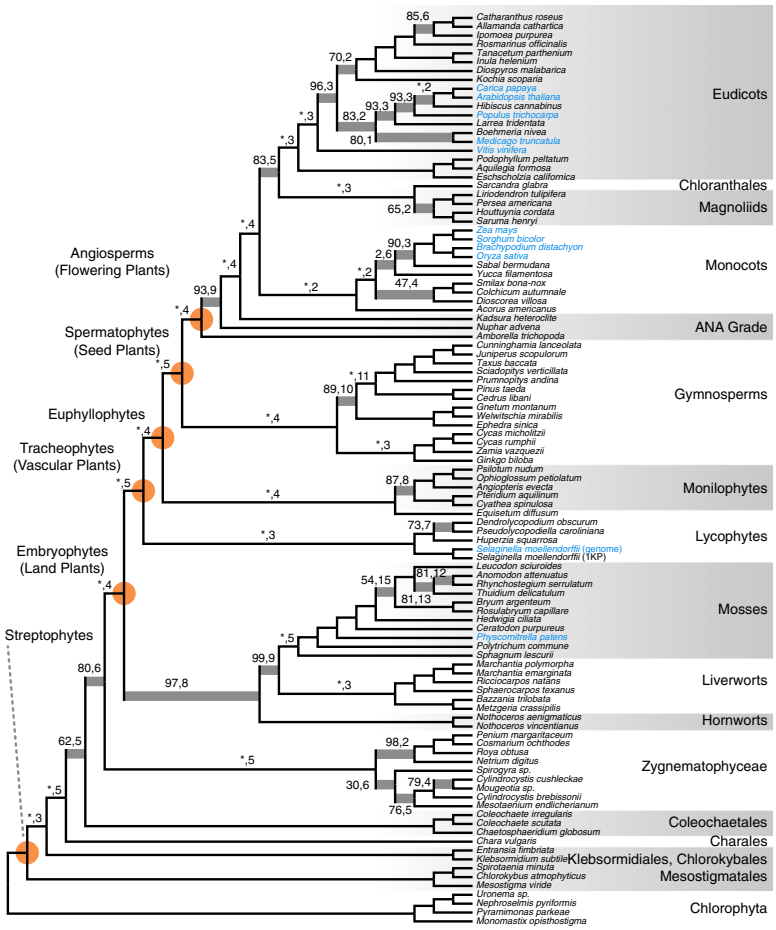

Fig. 3. Coalescent-based tree estimated from 424 gene trees estimated from first and second codon position alignments after removing genes with less than 50% taxon occupancy after gene fragments missing more than 66% of their sites were removed. Bootstrap values for the ASTRAL analysis are shown on branches (* for 100%) along with the percentages of gene trees exhibiting significant conflict (bootstrap support >75%) with nodes discussed in the text. See Figs. S8 and S9 for bootstrap values and percentages of conflicting gene trees for all nodes in analyses of first and second codon positions and amino acids, respectively. Most nodes exhibited 100% bootstrap support in the ASTRAL analysis, but gray branches highlight nodes with less than 100% support. Gene models from sequenced genomes are indicated by blue text.

Bryophyte Relationships. Whereas the monophyly of each bryophyte lineage—Bryophyta (mosses), Anthocerotophyta (hornworts), and Marchantiophyta (liverworts)—is strongly supported here (Figs. 2–4), most of our results reject the current, widely accepted hypothesis that liverworts are sister to all other land plants (38, 42, 71). Furthermore, the widely accepted view that liverworts, mosses, and hornworts are, respectively, successive sister groups to vascular plants (25, 42, 43)—which is strongly supported by parsimony mapping of mitochondrial intron gains (42) and recent mitochondrial phylogenomic analyses (72)—is not recovered in any of our analyses.

Previous analyses of protein-coding genes extracted from whole plastome sequences had suggested that the three bryophyte divisions (Bryophyta, Anthocerotophyta, and Marchantiophyta) form a clade (41, 73; but see ref. 71). Bryophytes are resolved as monophyletic in several analyses here, including 3 of 12 amino acid supermatrix analyses and all ASTRAL analyses based on either amino acid data or in-frame nucleotide data without the inclusion of third positions (Figs. 2–4). Supertree analyses of ML gene trees estimated from first and second codon position alignments and amino acids also favored this hypothesis (Fig. S4). For all analyses in which the three bryophyte lineages were resolved as a clade, mosses and liverworts formed a clade. In cases where a bryophyte clade was not recovered, our analyses generally recovered a clade with mosses and liverworts as sister to the tracheophytes, with the hornworts sister to all other (nonhornwort) land plants (Figs. 2–4), which is consistent with some previously published multigene analyses (40). Recent analyses of complete plastomes (8) and a PhyloBayes (59) analysis of amino acids under the CAT+GTR+Gamma substitution model (Fig. 4) (FAA.604-genes.trimExtensively.phylobayes.CATGTR) suggest a similar result, but with hornworts rather than a moss+liverwort clade sister to vascular plants. Independent chains in some PhyloBayes analyses (CAT+GTR+ Gamma analysis of first and second codon positions and CAT analysis of amino acids) recovered mosses, liverworts, and hornworts in a grade as successive sister clades to the tracheophytes (alignments and trees available in iPlant data store; mirrors.iplantcollaborative.org/onekp_pilot).

ML analyses were performed with the Gamma model of rate heterogeneity. Full GTRGAMMA and the per site rate (PSR) approximation of Gamma implemented in RAxML (74) produced nearly identical trees (Fig. 4). In addition, we performed partitioned analyses that assigned different amino acid substitution matrices or GTR matrices (for DNA) to different partitions of the data (see *Materials and Methods* for details). The CAT model implemented in PhyloBayes uses a Dirichlet process to model among-site variation in equilibrium state frequencies (75). The additional complexity of the CAT+GTR+Gamma model relative to the GTR+Gamma model may more closely match true variation in the substitution process (26, 75–77), but the difference in trees

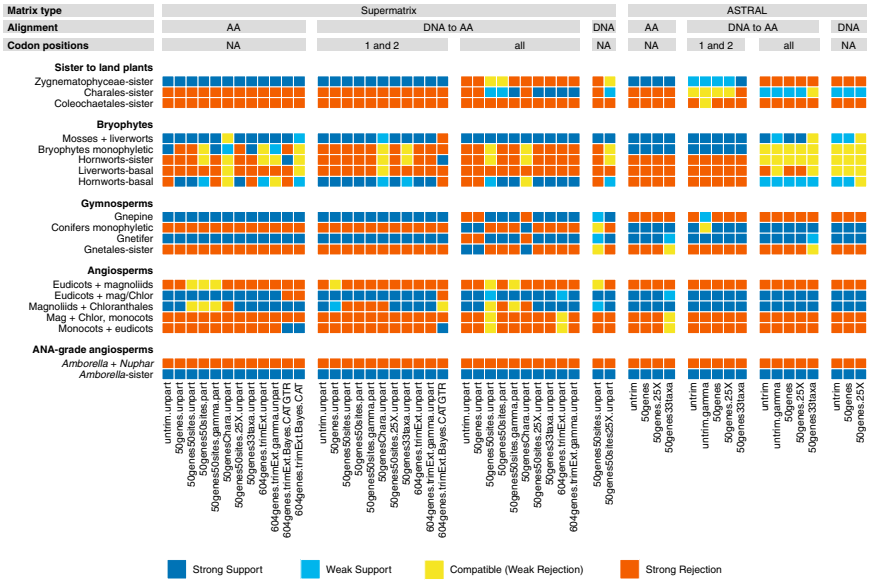

Fig. 4. Summary of support for hypotheses of land plant relationships across 52 supermatrix and coalescent-based analyses including permutations of the full data matrix (Table S2). Occupancy-based gene trimming was carried out by removing genes for which >50% the full taxon set were not included in the alignment. Site trimming removed columns in the alignment for which >50% of the full taxon set were represented by gap characters. Long-branch trimming was performed on gene trees when a terminal branch was 25-times longer than the median branch length. More stringent, blast-based removal of sequences identified as possible contaminants resulted in a set of 604 gene families (see *SI Materials and Methods* for stringent filtering strategy). Supermatrix analyses were done with and without partitioning of genes into model parameter classes. Filtering and analysis strategies are indicated below each column and include combinations of: (*i*) untrim: untrimmed/unfiltered data; (*ii*) unpart: no data partitions; (*iii*) 50genes: occupancy-based gene trimming at 50%; (*iv*) 50sites: occupancy-based site trimming at 50%; (*v*) gamma: full Gamma (vs. PSR approximation of Gamma); (*vi*) *Chara*: gene trimming to exclude genes not present in *Chara vulgaris*; (*vii*) 25X: long branch trimming; (*viii*) 33taxa: sequences with more than 66% gaps in gene alignments removed; (*ix*) 604genes.trimEXT: aggressive BLAST-based and long-branch filtering of sequence assemblies for each taxon followed by GBLOCKS filtering of sites (*SI Materials and Methods*). Strong support refers to bootstrap values above 75% for a clade containing the specified taxa. All trees and alignments are available in iPlant's data store (mirrors.iplantcollaborative.org/onekp_pilot).

estimated using CAT+GTR+Gamma models on nucleotide (first and second codon positions) and amino acid alignments suggests that this model may still be too simple for concatenated alignments relative to the true gene coalescence and substitution processes (see also ref. 72). The placement of hornworts and a moss+liverwort clade as successively sister to vascular plants is consistent with analyses based on morphological and developmental characters (78, 79), including dextral sperm in hornworts rather than sinistral sperm, as in all other land plants, and the retention of the pyrenoid, a plastid structure that is the site of RUBISCO localization, shared by hornworts and streptophytic algae (reviewed in ref. 80). The possibility that some of these trait mappings are the product of evolutionary convergence should also be considered, and seems likely in the case of the pyrenoid (81). The significance of other morphological similarities is also not yet clear. For example, the development of gametangia in hornworts resembles antheridial (44, 82) and archegonial (82) development in monilophytes, whereas those of the liverworts and mosses are autapomorphic, suggesting a closer relationship between hornworts and vascular plants. A comparison can also be made with respect to the development of the embryo and the young sporophyte. The hornwort embryo and sporophyte have no apical growth at any stage, but rather exhibit an intercalary meristem. In contrast, mosses and monilophytes have apical growth on both ends of the sporophyte, although basal apical growth is ephemeral in the former. The possibility of multiple origins of the multicellular sporophyte in land plants can therefore be considered (83): once with intercalary growth, as in the hornworts, and once with apical growth, as in mosses and tracheophytes (liverworts have neither intercalary nor apical growth). Ultimately, this finding underscores the difficulty in placing hornworts—or bryophytes in general—within the phylogeny of land plants based on current evidence from morphology alone.

In summary, three primary hypotheses emerge from our analyses with respect to the resolution of the earliest branching events in land plant phylogeny: (*i*) (hornworts, ((liverworts, mosses), vascular plants)) supported in most ML analyses of nucleotide and amino acid supermatrices; (*ii*) [(liverworts, mosses), (hornworts, vascular plants)], supported by the PhyloBayes analysis of amino acids; and (*iii*) [(hornworts, [mosses, liverworts]), vascular plants], supported by supertree and ASTRAL analyses of amino acids and first and second codon positions and some amino acid supermatrix analyses. However, we cannot dismiss alternative hypotheses recovered by some of our analyses, including [mosses (liverworts [hornworts, vascular plants])], which is supported by the PhyloBayes analysis of first and second codon positions (Fig. 4). Caution should be taken in rejecting any of these hypotheses given the sparse sampling, especially for the hornworts.

Monilophyte and Lycophyte Relationships. Phylogenetic analyses of multigene (generally plastid) datasets (84–87) have consistently resolved the lycophytes and monilophytes as successive sister lineages to the seed plants, with the euphyllophytes comprising the seed-free monilophytes (ferns) and seed-bearing spermatophytes. Aside from the clearly artifactual placement of *Selaginella* as sister to all other land plants in analyses including third codon positions (mirrors.iplantcollaborative.org/onekp_pilot), our results support this branching order (Figs. 2 and 3; other species trees at mirrors.iplantcollaborative.org/onekp_pilot). The placement of *Selaginella* has been problematic in previous analyses (49) and we interpret its misplacement in several analyses here as a consequence of GC content at the third codon position, which is more similar to streptophyte algae than to embryophytes (Fig. S2).

Monilophytes comprise (88) Psilotales (represented here by *Psilotum*), Ophioglossales (*Ophioglossum*), Equisetales (*Equisetum*), Marattiales (*Angiopteris*), and the leptosporangiate ferns (*Cyathea* and *Pteridium*). Here, Marattiales are consistently resolved as sister to a clade comprising Ophioglossales and Psilotales. Although the results here are inconsistent with previous analyses (84–86, 88), the resolution of the backbone phylogeny of ferns has been problematic (86), and we therefore interpret our results tentatively. Within the monilophytes, the placement of *Equisetum* varies (species trees at mirrors.iplantcollaborative.org/onekp_pilot) among analyses, as expected given the instability in the placement of *Equisetum* in many previous analyses (49, 84–86, 88). The number of loci used here to resolve the backbone of the streptophyte phylogeny is unprecedented; although extinction may play a significant role in the difficulty of reconstructing these relationships, analyses that include additional taxon sampling may contribute to a more robust set of relationships within land plant clades, particularly among fern lineages.

Gymnosperm Relationships. A well-supported seed plant clade was found, composed of strongly supported angiosperm and gymnosperm clades in all analyses (Figs. 2 and 3). Analyses varied, however, in the resolution of relationships among extant gymnosperms (Fig. 4). Whereas supermatrix analyses of alignments including all three codon positions placed Gnetales as sister to all other extant gymnosperm lineages (a hypothesis seen in refs. 52, 89, and 90), analyses of amino acids and of first and second codon positions placed Gnetales as sister to the Coniferales ["Gnetifer" hypothesis (Fig. 1); supertree and ASTRAL results (Figs. 1 and 3, and Fig. S4)] (91) or sister to the Pinaceae, nested within the Coniferales ["Gnepine" hypothesis (Figs. 1 and 2)] (26, 29, 35, 92–95). All but one of the ASTRAL and supertree analyses supported the Gnetifer hypothesis. Although most individual gene trees do not exhibit high bootstrap values, there were more gene trees exhibiting well-supported conflict with the Gnepine clade (Fig. 2) than the conifer clade (Fig. 3), and slightly more gene trees provide well-supported phylogenetic signal for the monophyly of Coniferales over a Gnetales+Pinaceae clade (Fig. S3). However, placement of Gnetales as sister to Pinaceae in most supermatrix analyses is consistent with previously published analyses of concatenated gene alignments that explicitly aimed to reduce long-branch attraction artifacts by filtering the most rapidly evolving sites (93, 95, 96) or implementing the CAT model discussed above (26, 75). In any case, these results are consistent with rapid diversification among the Gnetales and two conifer lineages; a scenario under which incomplete lineage sorting may mislead supermatrix analyses.

Angiosperm Relationships. Darwin famously referred to the rapid diversification of flowering plant lineages in the early history of angiosperms as an "abominable mystery" (97) and resolution of the earliest branching events remains controversial. Since publication of a series of landmark papers that identified *Amborella trichopoda*, Nymphaeales, and Austrobaileyales as successive sister lineages relative to all other extant angiosperms (98–101), all analyses performed with rich taxon sampling have supported *A. trichopoda* (7, 12, 14, 101–105) or a Nymphaeales+*A. trichopoda* clade (12, 106, 107) as sister to all other extant angiosperm lineages. All of our analyses placed *A. trichopoda* as sister to all other angiosperms (Figs. 2–4; species trees at mirrors.iplantcollaborative.org/onekp_pilot), with the Nymphaeales (represented by *Nuphar advena*) and Austrobaileyales (represented by *Kadsura heteroclite*) as successive sister lineages (i.e., the Amborellales, Nymphaeales, and Austrobaileyales or ANA grade) to the remaining angiosperms. This result is consistent with recent phylogenomic analyses of nuclear genes with many fewer sampled angiosperms and genes (35, 52, 89, 90, 107) and most earlier publications (98–101).

Resolving relationships among eudicots, monocots, and magnoliids has been a recalcitrant problem. All possible relationships among these three clades have been reported in the literature, but most recent analyses of plastid genomes have reconstructed magnoliids+Chloranthales as sister to (monocots + (eudicot+ Ceratophyllaceae)) (7, 14, 102). Resolution of these major angiosperm lineages varied among analyses. Ceratophyllaceae were not included in our analysis, but the Phylobayes CAT+GTR analyses of amino acid supported a magnoliid+Chloranthales (represented by *Sarcandra glabra*) clade sister to eudicots+ monocots (Fig. 4) (7, 14, 102). In contrast, RAxML GTR+ Gamma supermatrix, supertree, and ASTRAL analyses of amino acid and nucleotide alignments placed monocots outside of a (magnoliid, Chloranthales, eudicot) clade (Figs. 2–4). The placement of *S. glabra* (Chloranthales) varied between supermatrix analyses performed with and without filtering of trees including extreme branches or BLAST-based approaches to the filtering of contaminants (Figs. 2 and 4) (see *Materials and Methods* for more details), but all supertree and ASTRAL analyses recovered *Sarcandra* as sister to the magnoliids (Figs. 3 and 4).

Relationships within the magnoliid, monocot, and eudicot clades are largely in line with previously published analyses (14, 108), with the exception of the placement of *Vitis* as sister to the rest of the core eudicots including rosids, asterids, and Caryophyllales (Figs. 2 and 3; trees at mirrors.iplantcollaborative.org/onekp_pilot). *Vitis* is a rosid, but its placement can be problematic when taxon sampling is poor (6).

Variation in relationships inferred by different methods of analysis may be suggestive of model misspecification or variation in gene histories, perhaps because of incomplete lineage sorting. Problems with model misspecification may be resolved with the development of richer evolutionary models and our ongoing work to increase taxon sampling. Increased taxon sampling can reduce the effects of long-branch artifacts that are exacerbated by overly simplistic models of character evolution (109).

Conclusions

We present here a large-scale, phylogenomic perspective to resolving the backbone phylogeny of land plants and their closest green algal sister groups using a larger taxon set and more nuclear genes than have previously been applied to this problem. Our results are consistent with recent, algae-centric analyses that report Zygnematophyceae as sister to land plants (27–29, 36). However, our analyses suggest that the consistently accepted branching order of bryophytes (successive sister groups of liverworts, mosses, hornworts) should be reconsidered. Our results are largely consistent with a clade comprising mosses and liverworts, which agrees with recent analyses of plastomes (8); this clade is either sister to tracheophytes, sister to a clade composed of hornworts and tracheophytes, or included in a clade comprising all three bryophyte lineages (i.e., monophyletic bryophytes). Increased sampling of hornworts may help resolve their position across all types of analyses. Within monilophytes, and inconsistent with previous analyses (84, 86), we consistently recovered Marattiales sister to Ophioglossales plus Psilotales. We recovered strong support for a clade including Gnetales and Coniferales but in contrast to many phylogenomic analyses of plastid genomes (26, 28, 36, 92–95) our supertree and coalescent-based ASTRAL analyses placed Gnetales as sister to the Coniferales. Although concordance is not perfect, our results are generally in agreement with recent analyses of whole plastome data (8).

Despite the large number of nuclear genes included in this study, some relationships (e.g., the position of hornworts) remain enigmatic, perhaps because of extinction and ancient radiation, highlighting the need to evaluate the sources of incongruent signal in large datasets. However, the strength of some relationships in the face of analytical permutations (e.g., Zygnematophyceae sister to embryophytes, liverworts sister to mosses, and *Amborella* sister to

the remaining angiosperms), and the robust support for relationships inconsistent with currently accepted hypotheses (e.g., mosses plus liverworts monophyletic), emphasize the value of large nuclear datasets for phylogenetic reconstruction.

Materials and Methods

Tissue Collection, RNA Extraction, and Sequencing. Plant tissue was collected for—and provided to the project by—individual collaborators of the 1KP consortium (Table S1 for details). RNAs were isolated and transcriptomes were sequenced using protocols described previously (110). Briefly, plant tissues were collected, RNA was extracted and purified using protocols appropriate to the sample (108), and Illumina libraries were prepared. In some cases, plant material was shipped to the core sequencing facilities at the Beijing Genomics Institute (BGI)-Shenzhen and BGI-Hong Kong, in other cases purified total RNA was shipped to the sequencing facility. Sequencing libraries were prepared with an insert size of ∼200 bp. Multiple samples were multiplexed on a single lane of either Illumina GAIIx or HiSeq 2000 systems, with each sample sequenced to an approximate depth of 2 Gb with paired-end (2 × 75- or 2 × 90-bp) reads. As part of the BGI's methodology, read pairs that failed a minimum quality threshold were not de-multiplexed, and were discarded.

Transcriptome Assembly. RNA-Seq reads were assembled using SOAPdenovo v1 (111). Assembly was carried out using default parameters, with the exception of the use of 29-mers in deBruijn graph construction. The associated GapCloser tool was run as a postprocessing step to complete the assembly. The identity of the resulting assemblies was verified and checked for contamination through blastn searches against a custom database of 18S ribosomal RNA sequences.

Gene Family Circumscription and Transcriptome Sorting. To sort assembled transcripts into gene families, we constructed an a priori set of gene families by clustering the protein sequences of 25 sequenced plant genomes using orthoMCL (54). Clusters were searched to identify gene families that were predominantly single copy; given the frequency with which genes duplicate, we did not remove gene families in which a single taxon was represented by more than one gene (with a maximum of four genes for that taxon). Each cluster was then aligned using MAFFT (112) and the alignment was then used to build a profile HMM (pHMM) using HMMER3 (55).

Transcriptome assemblies were translated into matching amino acid and coding sequences using a strategy modified from TransPipe (113). An initial BLAST (blastx) (114) against all plant National Center for Biotechnology Information RefSeq proteins identified the best hit, which was then used to generate a GeneWise (115) translation. The resulting protein sequences were used to query the 25-genome pHMMs using hmmscan (part of the HMMER3 suite). Bit-scores for matches with e-values better than $1.0e^{-10}$ were retained and a cumulative probability distribution for the bit-scores was assessed to identify one or more HMMs accounting for 95% of the distribution. Most transcripts were sorted into a single gene family for which the HMM match had a probability of 95% or greater, but some transcripts were sorted into (and retained in) two or more families when bit score probabilities were required from multiple HMMs in order to reach a 95% confidence level that the assembly was sorted to a correct gene family (i.e., orthoMCL cluster).

For each gene family inferred to be low-copy, all transcripts that were sorted into a gene family for a single taxon were aligned to the 25-genome alignment to assess whether the transcripts could be scaffolded into a single sequence. Following the alignment step, the reference genome sequences were removed from the alignment and a consensus sequence was created from the query sequences using custom Perl scripts (available through the iPlant Discovery Environment: mirrors.iplantcollaborative.org/onekp_pilot). The number of non-A, C, T, and G bases in the consensus sequence was used to assess whether overlapping transcripts were paralogs or perhaps divergent alleles. If the number of non-A, C, T, and G bases per number of overlapping bases was greater than 5%, it was inferred that a gene duplication may have occurred in that lineage. In these cases, a sequence for that taxon/gene combination was not used in subsequent phylogenetic analyses.

Phylogenetic Analyses. Our 852 gene family files were each aligned using SATé (57), both as amino acid and nucleotide, resulting in two distinct alignments per gene family. We also forced nucleotide sequences on the amino acid alignments using a custom Perl script to obtain codon-preserving alignments of nucleotide sequences. Gene trees were then reconstructed for each gene family using RAxML (58) with 200 replicates of bootstrapping (average bootstrap support was centered around 50% across different gene trees) (Fig. S4), and based on 10 different starting trees. For each gene family, we estimated four different gene trees based on: (*i*) amino acid alignments, (*ii*) DNA alignment, (*iii*) codon alignments (nucleotides forced to the amino acid alignment), and (*iv*) codon alignments with third-position removed. Nucleotide-based analyses were conducted using the GTR model; for amino acid analyses, we used a Perl script (publicly available on the RAxML website) to score an estimated tree topology using different models, and selected the model that gave the highest likelihood score (Fig. S5). The JTT and JTTF models (116) had the highest likelihood score for 65% of genes. For handling rate heterogeneity across sites, we used the Gamma model for the main analyses, but for further exploration of parameters, we used the PSR approximation to Gamma (74), which consists of searching using 20 rate categories, and scoring and selecting the best tree using the Gamma model.

For supermatrix analyses, we concatenated all 852 gene alignments (1,701,170 bp), and then created multiple filtered datasets by: (*i*) removing genes that included 50% of taxa or less (674 genes and 1,414,611 bp left); (*ii*) removing sites with more than 50% missing characters (436,077 bp remaining); (*iii*) removing genes that did not include a sequence from *Chara* (to test whether its placement was an artifact of poor gene sampling) or those that had 50% of taxa or less (282 gens and 575,339 bp remaining); (*iv*) removing taxa from individual genes when they were on branches at least 25-times longer than the median branch length (possibly suggesting contamination) for that gene and then removing sites with at least 50% gaps (final alignment 429,722 bp); and finally (*v*) an extensive trimming of sequences using a blastp-based and branch-length–based approach as the most stringent filter for possible contamination and GBLOCKS to remove poorly aligned positions. This most stringent filter resulted in the removal of 248 gene families (604 genes and 386,883 bp retained; alignments at mirrors.iplantcollaborative.org/onekp_pilot). Note that after filtering taxa on long branches, new gene trees were estimated for genes that had at least one sequence removed (between 180 and 273 genes for different datasets). These filtered supermatrix datasets were created for amino acid, codon (nucleotides forced to the amino acid alignment), and nucleotide alignments. In addition, we created a set of datasets where the third codon position was removed.

ML supermatrix analyses were performed using RAxML v7.3 (43). In all nucleotide analyses, the GTR model was used. Because JTT and JTTF models were selected as the best model for a majority of our gene families, we used JTTF in our unpartitioned RAxML analyses. Similar to gene trees, the Gamma model of rate heterogeneity across sites was used for the main analyses, and the PSR approximation for the exploratory analyses. Finally, we performed partitioned RAxML analyses to better handle rate heterogeneity across genes. For codon alignments, we used the *K*-means clustering algorithm (117) to partition the data into 15 clusters of genes based on the GTR rate matrices calculated during gene tree estimation. We empirically observed that $k = 15$ accounts for most variation, while avoiding partitions that are too small. For amino acid alignments, the model selected for each gene family in gene tree estimation process was used to group loci together into 11 partitions, each defined by one substitution matrix. Each RAxML supermatrix analysis used 10 different MP trees as initial starting trees; the resulting RAxML tree with the best final ML score was selected as the final tree. Support was inferred for branches on the final tree from 100 bootstrap replicates.

The extensively trimmed amino acid and nucleotide supermatrices were analyzed with the site-heterogeneous CAT+Gamma model using PhyloBayes MPI (118). For the amino acid and nucleotide alignments, the CATGTR+Gamma model, which is consistently a better fit to the data than the CAT+Gamma model and any site-homogeneous models (76, 77), was also used. However, because of a high computational burden, perfect convergence of the two chains was not reached. Although the chains reached a plateau for all monitored values (e.g., likelihood or number of profiles), the topology was not identical for the two independent chains; however, the differences were limited to clades within angiosperms with very short internal branches. Nevertheless, the topologies recovered by the three models are almost identical to that in Figs. 2 and 3. The most significant differences are: (*i*) [hornworts,([liverworts,mosses],tracheophytes)] versus [mosses, (liverworts, [hornworts,tracheophytes])] (AA-CAT) or [(mosses, liverworts),(hornworts, tracheophytes)] (AA-CATGTR and NT-CATGTR), (*ii*) monocots sister to eudicots+magnoliids versus sister to eudicots, and (*iii*) cycadales sister to *Ginkgo* versus sister to all remaining gymnosperms (AA-CATGTR and NT-CATGTR).

Coalescent-based analyses were run using ASTRAL (61) and the multilocus bootstrapping procedure (119) was used to draw support values. ASTRAL estimates species trees from unrooted gene trees as input, and maximizes the number of quartet trees shared between the gene trees and the species tree. ASTRAL has been shown to be statistically consistent under the multispecies coalescent model [using results from Allman et al. (120) and Degnan (121) that

show four-taxon species trees do not have anomaly zones], and yields better accuracy than other coalescent-based methods in simulated studies (61).

ASTRAL runs were performed based on four types of input: (*i*) all gene trees, (*ii*) on only gene trees with more than 50% of taxa, (*iii*) on gene trees estimated after removing fragmentary data (i.e., sequences with more than 66% gaps), and (*iv*) on gene trees estimated after taxa on long branches were removed. The filtering of fragmentary data were in particular important for accurate gene tree estimation, because fragmentary sequences can negatively impact the accuracy of gene trees and hence the species tree (inclusion of fragmentary data does not have the same kind of impact on the concatenation analyses).

The multilocus bootstrapping was performed as follows. First, a main ASTRAL tree was estimated with ML gene trees as input. We then created 200 replicate input datasets, using 200 bootstrap replicates available for each gene (by randomly associating replicates from different genes together). On each of these 200 replicates, we estimated an ASTRAL tree, and we used these to infer support on the main tree. Conflict between specific branches in the species tree and gene trees was calculated by finding the percentage of gene trees that were incompatible with a given branch in the species tree after collapsing branches with support below 75%.

In addition to ASTRAL, we also performed supertree analyses using Superfine-MRP (60), with TNT (122) used for the MRP step. The supertree analyses used the same multilocus bootstrapping (119) procedure that was used for ASTRAL. More details about phylogenetic reconstruction are available in *SI Materials and Methods* and Dataset S1. A compete description of resources associated with the 1KP data has been published (123).

ACKNOWLEDGMENTS. This work was supported largely by funding from the Alberta Ministry of Innovation and Advanced Education (G.K.-S.W.), Alberta Innovates Technology Futures (G.K.-S.W.), Innovates Centres of Research Excellence (G.K.-S.W.), Musea Ventures (G.K.-S.W.), and BGI-Shenzhen for The 1000 Plants (1KP) initiative (G.K.-S.W.); computation support was provided by the China National GeneBank (CNGB) and the Texas Advanced Computing Center (TACC); significant support, including personnel, computational resources, and data hosting, was also provided by the iPlant Collaborative as funded by the National Science Foundation (DBI-1265383) and National Science Foundation Grants IOS 0922742 (to C.W.d., P.S.S., D.E.S., and J.L.-M.), IOS 0922738 (to D.W.S.), DEB 0830009 (to J.L.-M., C.W.d., S.W.G., and D.W.S.), EF-0629817 (to S. Mathews, S.W.G., and D.W.S.), DEB 0733029 (to T.W. and J.L.-M.), and DBI 1062335 (to T.W.), a National Institutes of Health Grant 1R01DA025197 (to T.M.K., C.W.d., and J.L.-M.), and a Natural Sciences and Engineering Research Council of Canada Discovery grant (to S.W.G.).

1. Kenrick P, Crane PR (1997) The origin and early evolution of plants on land. *Nature* 389(6646):33–39.
2. Rubinstein CV, Gerrienne P, de la Puente GS, Astini RA, Steemans P (2010) Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). *New Phytol* 188(2):365–369.
3. Steemans P, et al. (2009) Origin and radiation of the earliest vascular land plants. *Science* 324(5925):353.
4. Wellman CH, Osterloff PL, Mohiuddin U (2003) Fragments of the earliest land plants. *Nature* 425(6955):282–285.
5. Lenton TM, Crouch M, Johnson M, Pires N, Dolan L (2012) First plants cooled the Ordovician. *Nat Geosci* 5(2):86–89.
6. Jansen RK, et al. (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: Effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol Biol* 6:32.
7. Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc Natl Acad Sci USA* 104(49):19363–19368.
8. Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG (2014) From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol Biol* 14:23.
9. Bremer B, et al. (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* 161(2):105–121.
10. Bremer B, et al. (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot J Linn Soc* 141(4):399–436.
11. Bremer K, et al. (1998) An ordinal classification for the families of flowering plants. *Ann Mo Bot Gard* 85(4):531–553.
12. Qiu YL, et al. (2010) Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *J Syst Evol* 48(6):391–425.
13. Barkman TJ, et al. (2007) Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. *BMC Evol Biol* 7:248.
14. Soltis DE, et al. (2011) Angiosperm phylogeny: 17 genes, 640 taxa. *Am J Bot* 98(4): 704–730.
15. Burleigh JG, Hilu KW, Soltis DE (2009) Inferring phylogenies with incomplete data sets: A 5-gene, 567-taxon analysis of angiosperms. *BMC Evol Biol* 9:61.
16. Devereux R, Loeblich AR, 3rd, Fox GE (1990) Higher plant origins and the phylogeny of green algae. *J Mol Evol* 31(1):18–24.
17. Manhart JR (1994) Phylogenetic analysis of green plant *rbcL* sequences. *Mol Phylogenet Evol* 3(2):114–127.
18. Manhart JR, Palmer JD (1990) The gain of two chloroplast tRNA introns marks the green algal ancestors of land plants. *Nature* 345(6272):268–270.
19. Melkonian M, Surek B (1995) Phylogeny of the Chlorophyta—Congruence between ultrastructural and molecular evidence. *Bulletin De La Societe Zoologique De France-Evolution Et Zoologie* 120(2):191–208.
20. Surek B, Beemelmanns U, Melkonian M, Bhattacharya D (1994) Ribosomal-RNA sequence comparisons demonstrate an evolutionary relationship between Zygnematales and Charophytes. *Plant Syst Evol* 191(3-4):171–181.
21. Bremer K (1985) Summary of green plant phylogeny and classification. *Cladistics* 1(4):369–385.
22. Kenrick P, Crane PR (1997) *The Origin and Early Diversification of Land Plants: A Cladistic Study* (Smithsonian Institution Press, Washington, DC), pp xi, 441 pp.
23. Lemieux C, Otis C, Turmel M (2007) A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biol* 5:2.
24. Mishler BD, Churchill SP (1985) Transition to a land flora: Phylogenetic relationships of the green algae and bryophytes. *Cladistics* 1(4):305–328.
25. Qiu YL, et al. (2006) The deepest divergences in land plants inferred from phylogenomic evidence. *Proc Natl Acad Sci USA* 103(42):15511–15516.
26. Laurin-Lemay S, Brinkmann H, Philippe H (2012) Origin of land plants revisited in the light of sequence contamination and missing data. *Curr Biol* 22(15):R593–R594.
27. Timme RE, Bachvaroff TR, Delwiche CF (2012) Broad phylogenomic sampling and the sister lineage of land plants. *PLoS ONE* 7(1):e29696.
28. Turmel M, Otis C, Lemieux C (2006) The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. *Mol Biol Evol* 23(6):1324–1338.
29. Wodniok S, et al. (2011) Origin of land plants: Do conjugating green algae hold the key? *BMC Evol Biol* 11:104.
30. Zhong B, Liu L, Yan Z, Penny D (2013) Origin of land plants using the multispecies coalescent model. *Trends Plant Sci* 18(9):492–495.
31. Fowke LC, Pickett-Heaps JD (1969) Cell division in *Spirogyra*. II. Cytokinesis. *J Phycol* 5(4):273–281.
32. Galway ME, Hardham AR (1991) Immunofluorescent localization of microtubules throughout the cell-cycle in the green-alga *Mougeotia* (Zygnemataceae). *Am J Bot* 78(4):451–461.
33. Pickett-Heaps JD, Wetherbee R (1987) Spindle function in the green-alga *Mougeotia*—Absence of anaphase a correlates with postmitotic nuclear migration. *Cell Motil Cytoskeleton* 7(1):68–77.
34. Karol KG, McCourt RM, Cimino MT, Delwiche CF (2001) The closest living relatives of land plants. *Science* 294(5550):2351–2353.
35. Finet C, Timme RE, Delwiche CF, Marlétaz F (2010) Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr Biol* 20(24):2217–2222.
36. Civán P, Foster PG, Embley MT, Séneca A, Cox CJ (2014) Analyses of charophyte chloroplast genomes help characterize the ancestral chloroplast genome of land plants. *Genome Biol Evol* 6(4):897–911.
37. Turmel M, Pombert JF, Charlebois P, Otis C, Lemieux C (2007) The green algal ancestry of land plants as revealed by the chloroplast genome. *Int J Plant Sci* 168(5):679–689.
38. Chang Y, Graham SW (2011) Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. *Am J Bot* 98(5):839–849.
39. Shaw AJ, Szövényi P, Shaw B (2011) Bryophyte diversity and evolution: Windows into the early evolution of land plants. *Am J Bot* 98(3):352–369.
40. Nickrent DL, Parkinson CL, Palmer JD, Duff RJ (2000) Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol Biol Evol* 17(12):1885–1895.
41. Nishiyama T, et al. (2004) Chloroplast phylogeny indicates that bryophytes are monophyletic. *Mol Biol Evol* 21(10):1813–1819.
42. Qiu YL, Cho Y, Cox JC, Palmer JD (1998) The gain of three mitochondrial introns identifies liverworts as the earliest land plants. *Nature* 394(6694):671–674.
43. Qiu YL, et al. (2007) A nonflowering land plant phylogeny inferred from nucleotide sequences of seven chloroplast, mitochondrial, and nuclear genes. *Int J Plant Sci* 168(5):691–708.
44. Renzaglia KS, Nickrent DL, Garbary DJ, Garbary DJ; Duff RJT (2000) Vegetative and reproductive innovations of early land plants: Implications for a unified phylogeny. *Philos Trans R Soc Lond B Biol Sci* 355(1398):769–793.
45. Ligrone R, Duckett JG, Renzaglia KS (2012) Major transitions in the evolution of early land plants: A bryological perspective. *Ann Bot (Lond)* 109(5):851–871.
46. Remy W, Gensel PG, Hass H (1993) The gametophyte generation of some early Devonian land plants. *Int J Plant Sci* 154(1):35–58.
47. Cox CJ, Li B, Foster PG, Embley TM, Civán P (2014) Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Syst Biol* 63(2):272–279.
48. Groth-Malonek M, Pruchner D, Grewe F, Knoop V (2005) Ancestors of trans-splicing mitochondrial introns support serial sister group relationships of hornworts and mosses with vascular plants. *Mol Biol Evol* 22(1):117–125.
49. Karol KG, et al. (2010) Complete plastome sequences of *Equisetum arvense* and *Isoetes flaccida*: Implications for phylogeny and plastid genome evolution of early land plant lineages. *BMC Evol Biol* 10:321.
50. Mathews S (2009) Phylogenetic relationships among seed plants: Persistent questions and the limits of molecular data. *Am J Bot* 96(1):228–236.

51. Burleigh JG, et al. (2011) Genome-scale phylogenetics: Inferring the plant tree of life from 18,896 gene trees. *Syst Biol* 60(2):117–125.
52. Lee EK, et al. (2011) A functional phylogenomic view of the seed plants. *PLoS Genet* 7(12):e1002411.
53. Timme RE, Delwiche CF (2011) Phylogenomic reconstruction of the Charophytes: A multilocus approach to resolving the phylogeny of plants' closest relatives. *J Phycol* 47(Suppl 1):16.
54. Li L, Stoeckert CJ, Jr, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13(9):2178–2189.
55. Eddy SR (2011) Accelerated profile HMM searches. *PLOS Comput Biol* 7(10): e1002195.
56. Duarte JM, et al. (2010) Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol Biol* 10:61.
57. Liu K, et al. (2012) SATe-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst Biol* 61(1):90–106.
58. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
59. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25(17): 2286–2288.
60. Swenson MS, Suri R, Linder CR, Warnow T (2012) SuperFine: Fast and accurate supertree estimation. *Syst Biol* 61(2):214–227.
61. Mirarab S, et al. (2014) ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* 30(17):i541–i548.
62. Gontcharov AA, Melkonian M (2004) Unusual position of the genus *Spirotaenia* (Zygnematophyceae) among streptophytes revealed by SSU rDNA and *rbcL* sequence comparisons. *Phycologia* 43(1):105–113.
63. Maddison WP (1997) Gene trees in species trees. *Syst Biol* 46(3):523–536.
64. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol* 24(6):332–340.
65. Lewis LA, McCourt RM (2004) Green algae and the origin of land plants. *Am J Bot* 91(10):1535–1556.
66. Sawitzky H, Grolig F (1995) Phragmoplast of the green alga *Spirogyra* is functionally distinct from the higher plant phragmoplast. *J Cell Biol* 130(6):1359–1371.
67. Bakker ME, Lokhorst GM (1987) Ultrastructure of mitosis and cytokinesis in *Zygnema-sp* (Zygnematales, Chlorophyta). *Protoplasma* 138(2-3):105–118.
68. Jürgens G (2005) Plant cytokinesis: Fission by fusion. *Trends Cell Biol* 15(5):277–283.
69. Pickett-Heaps JD (1969) The evoltuion of mitotic apparatus—An attempt at comparative ultrastructural cytology in dividing plant cells. *Cytobios* 1(3):257–280.
70. Chapman RL, Borkhsenious O, Brown RC, Henk MC, Waters DA (2001) Phragmoplast-mediated cytokinesis in Trentepohlia: Results of TEM and immunofluorescence cytochemistry. *Int J Syst Evol Microbiol* 51(Pt 3):759–765.
71. Gao L, Su YJ, Wang T (2010) Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. *J Syst Evol* 48(2):77–93.
72. Liu Y, Cox CJ, Wang W, Goffinet B (2014) Mitochondrial phylogenomics of early land plants: Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Syst Biol*, 10.1093/sysbio/syu049.
73. Goremykin VV, Hellwig FH (2005) Evidence for the most basal split in land plants dividing bryophyte and tracheophyte lineages. *Plant Syst Evol* 254(1-2):93–103.
74. Stamatakis A, Aberer AJ (2013) Novel parallelization schemes for large-scale likelihood-based phylogenetic inference. *Parallel and Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on May 20–24* (IEEE, Washington, DC), pp 1195–1204.
75. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21(6):1095–1109.
76. Philippe H, Roure B (2011) Difficult phylogenetic questions: More data, maybe; better methods, certainly. *BMC Biol* 9:91.
77. Roure B, Baurain D, Philippe H (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol Biol Evol* 30(1):197–214.
78. Garbary DJ, Renzaglia KS, Duckett JG (1993) The phylogeny of land plants—A cladistic-analysis based on male gametogenesis. *Plant Syst Evol* 188(3-4):237–269.
79. Renzaglia KS, Duckett JG (1991) Towards an understanding of the differences between the blepharoplasts of mosses and liverworts, and comparisons with hornworts, biflagellate lycopods and charophytes—A numerical-analysis. *New Phytol* 117(2):187–208.
80. Vaughn KC, et al. (1992) The anthocerote chloroplast—A review. *New Phytol* 120(2): 169–190.
81. Meyer M, Griffiths H (2013) Origins and diversity of eukaryotic CO2-concentrating mechanisms: Lessons for the future. *J Exp Bot* 64(3):769–786.
82. Smith GM (1955) *Cryptogamic Botany Vol. II: Bryophytes and Pteridophytes* (McGraw Hill, New York).
83. Philipson WR (1991) A new approach to the origins of vascular plants. *Botanische Jahrbucher* 113:443–460.
84. Grewe F, Guo W, Gubbels EA, Hansen AK, Mower JP (2013) Complete plastid genomes from *Ophioglossum californicum, Psilotum nudum*, and *Equisetum hyemale* reveal an ancestral land plant genome structure and resolve the position of Equisetales among monilophytes. *BMC Evol Biol* 13:8.
85. Pryer KM, et al. (2001) Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature* 409(6820):618–622.
86. Rai HS, Graham SW (2010) Utility of a large, multigene plastid data set in inferring higher-order relationships in ferns and relatives (monilophytes). *Am J Bot* 97(9): 1444–1456.
87. Wolf PG, et al. (2005) The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). *Gene* 350(2):117–128.
88. Pryer KM, et al. (2004) Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. *Am J Bot* 91(10):1582–1598.
89. Cibrián-Jaramillo A, et al. (2010) Using phylogenomic patterns and gene ontology to identify proteins of importance in plant evolution. *Genome Biol Evol* 2:225–239.
90. de la Torre-Bárcena JE, et al. (2009) The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. *PLoS ONE* 4(6): e5764.
91. Chaw SM, Zharkikh A, Sung HM, Lau TC, Li WH (1997) Molecular phylogeny of extant gymnosperms and seed plant evolution: Analysis of nuclear 18S rRNA sequences. *Mol Biol Evol* 14(1):56–68.
92. Bowe LM, Coat G, dePamphilis CW (2000) Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proc Natl Acad Sci USA* 97(8):4092–4097.
93. Burleigh JG, Mathews S (2004) Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. *Am J Bot* 91(10): 1599–1613.
94. Qiu YL, Palmer JD (1999) Phylogeny of early land plants: Insights from genes and genomes. *Trends Plant Sci* 4(1):26–30.
95. Zhong B, Yonezawa T, Zhong Y, Hasegawa M (2010) The position of gnetales among seed plants: Overcoming pitfalls of chloroplast phylogenomics. *Mol Biol Evol* 27(12): 2855–2863.
96. Zhong B, et al. (2011) Systematic error in seed plant phylogenomics. *Genome Biol Evol* 3:1340–1348.
97. Darwin C, Darwin F, Seward AC (1903) in *More Letters of Charles Darwin: A Record of his Work in a Series of Hitherto Unpublished Letters*, eds Darwin F, Seward AC (J. Murray, London).
98. Mathews S, Donoghue MJ (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286(5441):947–950.
99. Parkinson CL, Adams KL, Palmer JD (1999) Multigene analyses identify the three earliest lineages of extant flowering plants. *Curr Biol* 9(24):1485–1488.
100. Qiu YL, et al. (1999) The earliest angiosperms: Evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402(6760):404–407.
101. Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402(6760):402–404.
102. Jansen RK, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci USA* 104(49):19369–19374.
103. Stefanović S, Rice DW, Palmer JD (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol Biol* 4:35.
104. Graham SW, Iles WJD (2009) Different gymnosperm outgroups have (mostly) congruent signal regarding the root of flowering plant phylogeny. *Am J Bot* 96(1):216–227.
105. Soltis DE, et al. (2000) Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. *Bot J Linn Soc* 133(4):381–461.
106. Goremykin VV, et al. (2013) The evolutionary root of flowering plants. *Syst Biol* 62(1):50–61.
107. Xi Z, Liu L, Rest JS, Davis CC (2014) Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Syst Biol*, 10.1093/sysbio/syu055.
108. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci USA* 107(10):4623–4628.
109. Leebens-Mack J, et al. (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Mol Biol Evol* 22(10):1948–1963.
110. Johnson MTJ, et al. (2012) Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLoS ONE* 7(11):e50226.
111. Li R, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265–272.
112. Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. *Bioinformatics for DNA Sequence Analysis, Methods in Molecular Biology*, ed Posada D (Humana, Totowa), Vol 537, pp 39–64.
113. Barker MS, et al. (2010) EvoPipes.net: Bioinformatic tools for ecological and evolutionary genomics. *Evol Bioinform Online* 6:143–149.
114. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
115. Birney E, Clamp M, Durbin R (2004) GeneWise and genomewise. *Genome Res* 14(5): 988–995.
116. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8(3):275–282.
117. Hartigan JA, Wong MA (1979) Algorithm AS 136: A *K*-means clustering algorithm. *J R Stat Soc Ser C Appl Stat* 28(1):100–108.
118. Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* 62(4):611–615.
119. Seo T-K (2008) Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol Biol Evol* 25(5):960–971.
120. Allman ES, Degnan JH, Rhodes JA (2011) Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J Math Biol* 62(6):833–862.
121. Degnan JH (2013) Anomalous unrooted gene trees. *Syst Biol* 62(4):574–590.
122. Goloboff PA, Farris JS, Nixon KC (2008) TNT, a free program for phylogenetic analysis. *Cladistics* 24(5):774–786.
123. Matasci N, et al. (2014) Data access for the 1,000 Plants (1KP) project. *GigaScience* 3:17.