Supporting Information

Wickett et al. 10.1073/pnas.1323926111

SI Materials and Methods

Meta-Assembly of Illumina, 454, and Sanger Sequence Data. In addition to the Illumina sequences data used for the majority of taxa included in this study, 454 sequence data were used for several species and Sanger sequences were used for three species (Table S1). The 454 sequences were retrieved from the National Center for Biotechnology Information's Sequence Read Archive (NCBI SRA) database and assembled using MIRA (1) (sourceforge.net/projects/mira-assembler); and assembled Sanger sequences were retrieved from PlantGDB (2) (www.plantgdb.org/prj/ESTCluster). Meta-assemblies of 454, Sanger, and SOAP de novo (3) assemblies of the 1KP data were generated using CAP3 (4) (seq.cs.iastate.edu), setting minimum overlap and sequence identify parameters at 25 bp and 0.95, respectively. The resulting meta-assemblies were posted along with all of the 1KP assemblies in the iPlant data store (mirrors.iplantcollaborative.org/onekp_pilot).

Stringent Filtering of the 852 Single-Copy Gene Set. All of the following analyses were performed at the amino acid level.

Removal of paralogous or contaminated sequences using blast. A database was constructed with formatdb from 15 proteomes downloaded from the NCBI website and divided in six reference species (*Arabidopsis thaliana*, *Oryza sativa*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Chlamydomonas reinhardtii*, *Micromonas pusilla*) representative of the green plants, and nine distantly related species roughly covering the remaining tree of life (*Cryptococcus neoformans*, *Aspergillus fumigatus*, *Botryotinia fuckeliana*, *Gibberella zeae*, *Homo sapiens*, *Phytophthora infestans*, *Capsaspora owczarzaki*, *Bacillus subtilis*, *Sinorhizobium meliloti*) used to pinpoint contaminations. For each sequence of each alignment file, a blast was performed by blastp, v2.2.20 (evalue $\geq 1e^{-10}$, no filter for low complexity, 100 hits recovered), to determine its closest relatives in the database. Only hits with coverage of at least 50% of the best hit are considered for determining paralogous or contaminated sequences.

A sequence is considered as a contamination if the best hit for at least one contaminant species has a log(e-value) greater than the log(e-value of the best hit over all reference species) plus 5.

For each reference species, a main protein (i.e., the protein that appears with the best hit for the most sequences of the alignment file) is considered as the orthologous protein. To avoid redundancy (including alternatively spiced forms) in the database, hits with an identity greater or equal to 95% against the main protein for a reference species are considered as similar to this main protein. Otherwise, the sequence is flagged as a paralogous sequence if the best hit is different of the main protein for at least four reference species or for all reference species if the sequence matches on less than four reference species.

To remove the genes that are a complex mix of paralogs, an alignment file is totally removed from the dataset if less than 90% of the sequences match on the main protein for at least four reference species. All of the flagged files were manually verified, because this step heavily relies on the quality of reference genomes, which is far from perfect. In particular, when the gene of interest is incorrectly annotated for several species (e.g., predicted as several independent genes), this criterion may be misled.

One-hundred forty-seven sequences were removed as contaminant and 634 as paralog; 22 genes were removed because of complex paralogy.

Removal of erroneous, paralogous or contaminated sequences using branch length. After this first step, frame-shifts were detected and removed using HmmCleaner (5). Briefly, this program removes the sequence blocks of a given species that do not fit well the HMM profile constructed from all of the species but this one. Then the alignments were filtered using GBLOCKS with parameters ($n1 = n/2 + 1$, $n2 = 0.55 \times n$, $n3 = 5$, $n4 = 5$, and $n5 = 2$, n being the number of species). These parameters are not very stringent to accommodate the large proportion of partial sequences. These filtered sequences were concatenated using Scafos and a tree was constructed from this supermatrix using Treefinder with a WAG+F+G model. This tree was then used as a reference topology.

For each gene, branch lengths for the reference topology were computed using Tree-puzzle with a WAG+F+G model (the missing species were first pruned from the reference topology). The branch lengths inferred for the single gene were compared with the ones of the reference topology following this protocol: the size of the two trees was normalized to 100 and then the branch leading to the leaves were compared. A species was considered as erroneous (because of paralogy, contamination, sequencing error, alignment error, and so forth) when the branch length was greater than five-times the branch of the reference tree. As a result, 113 sequences were removed according to this criterion.

Once these highly divergent sequences that can bias tree length normalization were removed, the analysis was performed again, with more stringent criteria. First, the sequences with less than 40 amino acids (before and especially after GBLOCKS) were discarded to avoid stochastic errors. Similarly, 13 genes with fewer than 40 positions after GBLOCKS were removed. Ninety-one sequences were removed because their branch length was greater than three-times the branch of the reference tree.

This stringent analysis led to a set of 815 single gene alignments. These sequences were concatenated using Scafos and only the genes for which at most 51 species (over 103) were missing have been considered. The supermatrix based on 604 genes following this criterion contained 128,956 positions, of which 105,236 were variable. There are 40% of missing data in this alignment. All of the steps performed at the amino acid level were reported on the nucleotide sequences. The third codon positions were removed after concatenation.

The amino acid and nucleotide supermatrices were analyzed with the site-heterogeneous CAT+G and CATGTR+G models, respectively, using phylobayes-mpi. Although the chains clearly reached a plateau for all of the monitored values (e.g., likelihood or number of profiles), the topology was not identical for the two independent chains. For the amino acids, the differences are (hornworts, ((liverworts, mosses), tracheophytes)) versus (mosses, (liverworts, (hornworts, tracheophytes))) and *Sarcandra* sister of eudicots or of the remaining magnoliids. For the nucleotides, the differences are *Sarcandra* sister of eudicots + the remaining magnoliids or of the remaining magnoliids and (Amborella, (Nuphar, remaining angiosperms)) versus (Nuphar, (Amborella, remaining angiosperms)).

For the amino acid alignments, the CATGTR+G model, which has always a better fit to the data than the CAT+G model, was also used, but because of high computational burden, a good convergence of the chains was not reached. Nevertheless the recovered topology by the two chains is almost identical to the one in Fig. 2, except that the hornworts are sister of tracheophytes.

1. Chevreux B, et al. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 14(6):1147–1159.
2. Duvick J, et al. (2008) PlantGDB: A resource for comparative plant genomics. *Nucleic Acids Res* 36(Database issue):D959–D965.
3. Li R, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265–272.
4. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res* 9(9): 868–877.
5. Amemiya CT, et al. (2013) The African coelacanth genome provides insights into tetrapod evolution. *Nature* 496(7445):311–316.

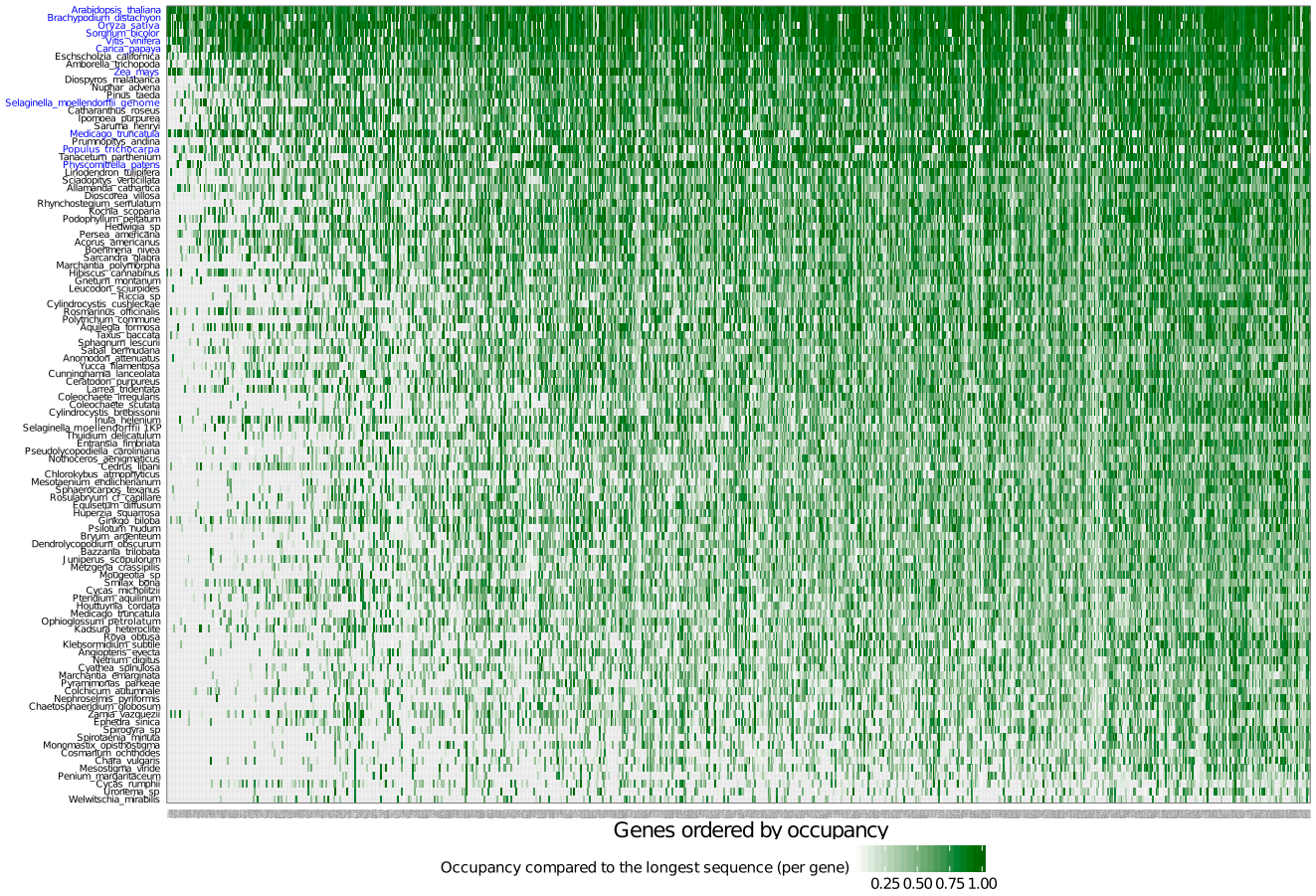

Fig. S1. Taxon occupancy for each taxon and gene. Each row corresponds to a taxon and each column corresponds to a single-copy gene family. Shades of green correspond to different levels of taxon occupancy, measured as the length of sequence from a taxon, compared with the length of the longest sequence from any taxa for that gene. Empty cells (i.e., those in white) indicate that a gene (column) was not sampled or retained after filtering for a taxon (row).

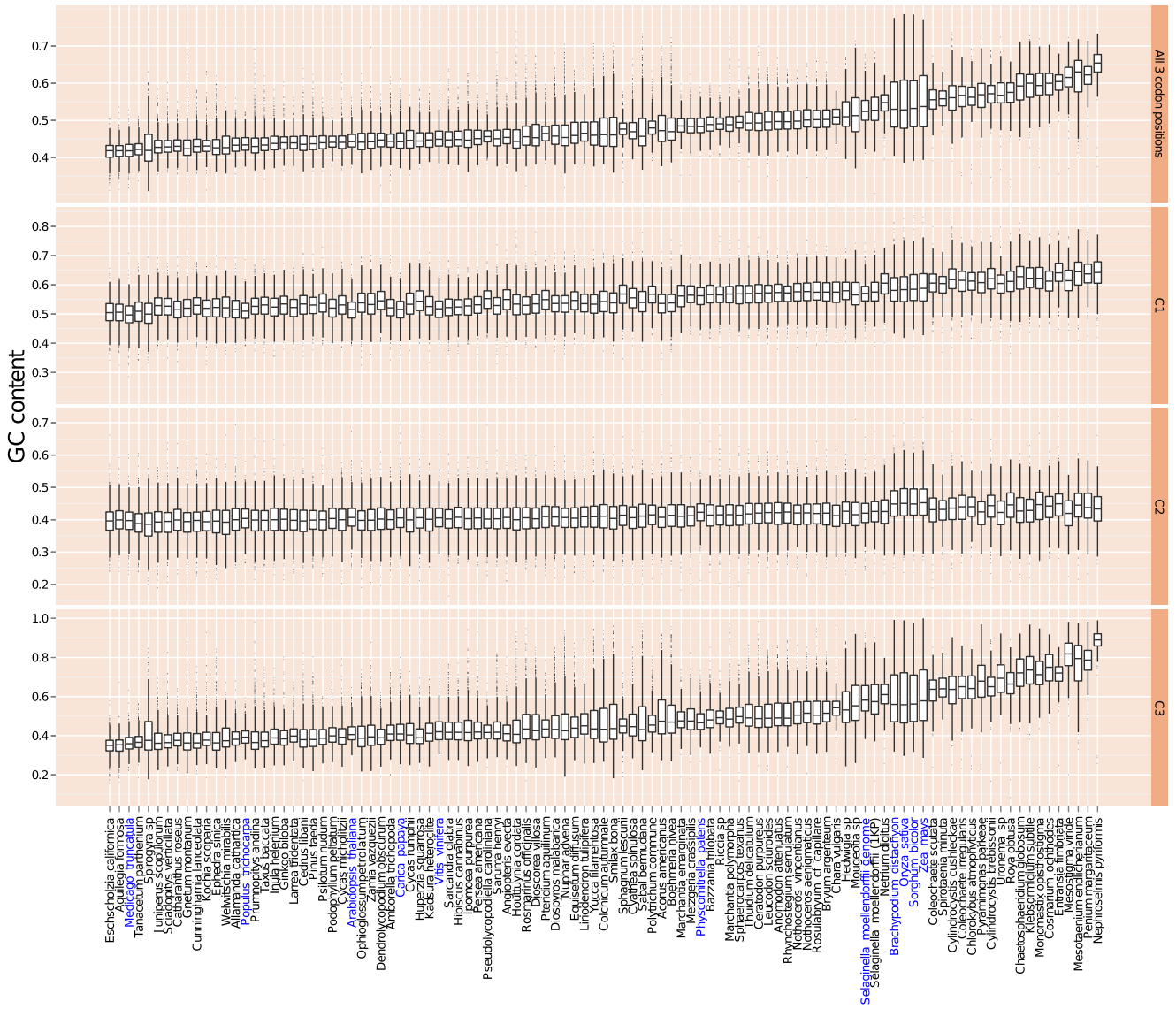

Fig. S2. Plots of GC content at first, second, and third codon positions indicate significant variation at the third position.

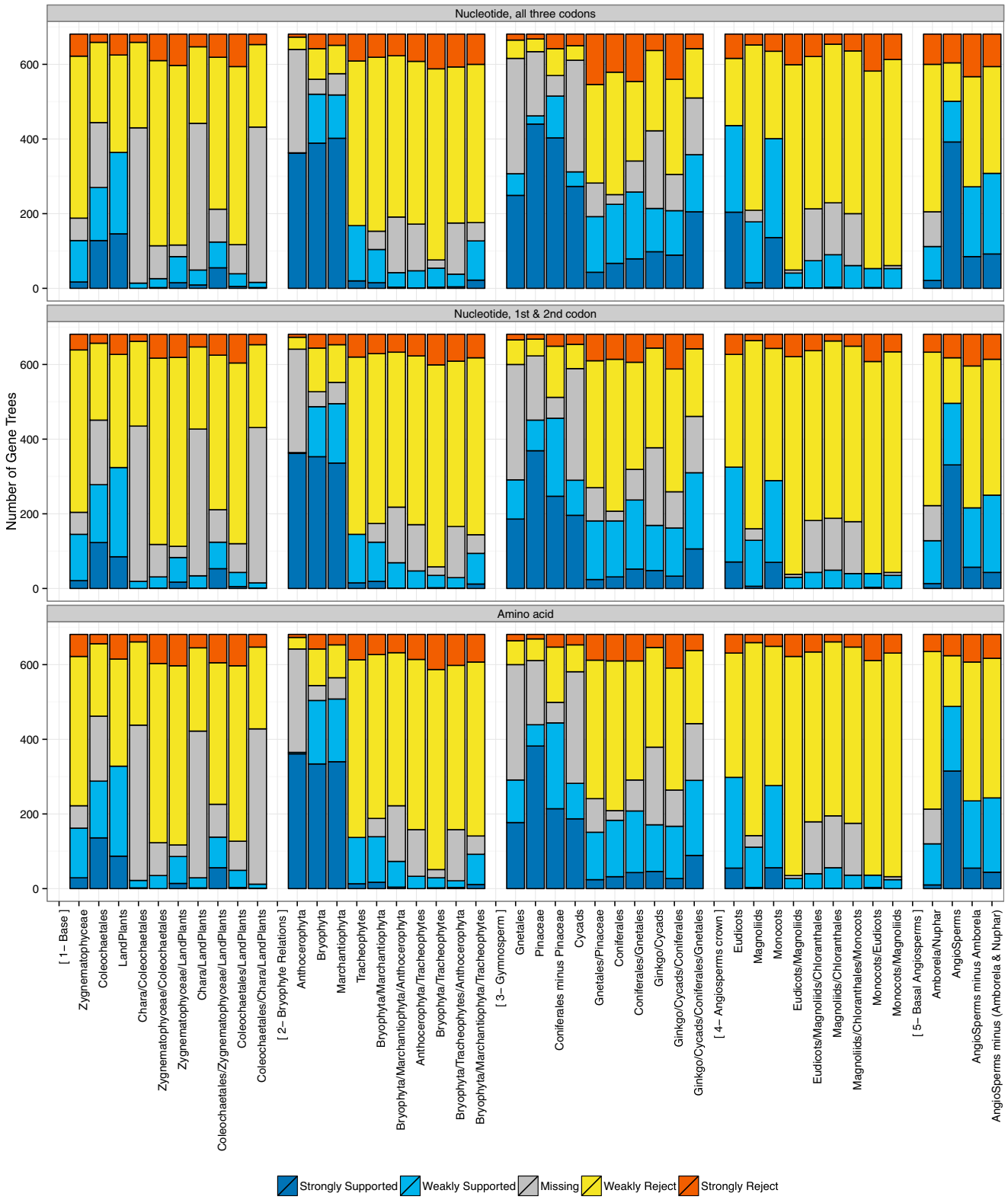

Fig. S3. Summary of support for alternative hypotheses in gene trees (Fig. 1). Each bar corresponds to a focal clade and represents all of the 681 single-copy gene trees that could be rooted using outgroups. Colors indicate the degree of support (dark blue or light blue) or conflict (cream or red) for the monophyly of the specified clade in each gene tree. Bootstrap support values greater than 75% are indicated as strong and 50–75% as weak support or rejection of specified clade. The monophyly of a clade was tested only if at least one taxon is present from each of listed component of that clade; gray portions of the bars represent trees where at least one component was entirely missing. For example, the monophyly of *Chara*/land plants is tested only for rooted gene trees with *Chara* and at least one land plant taxon.

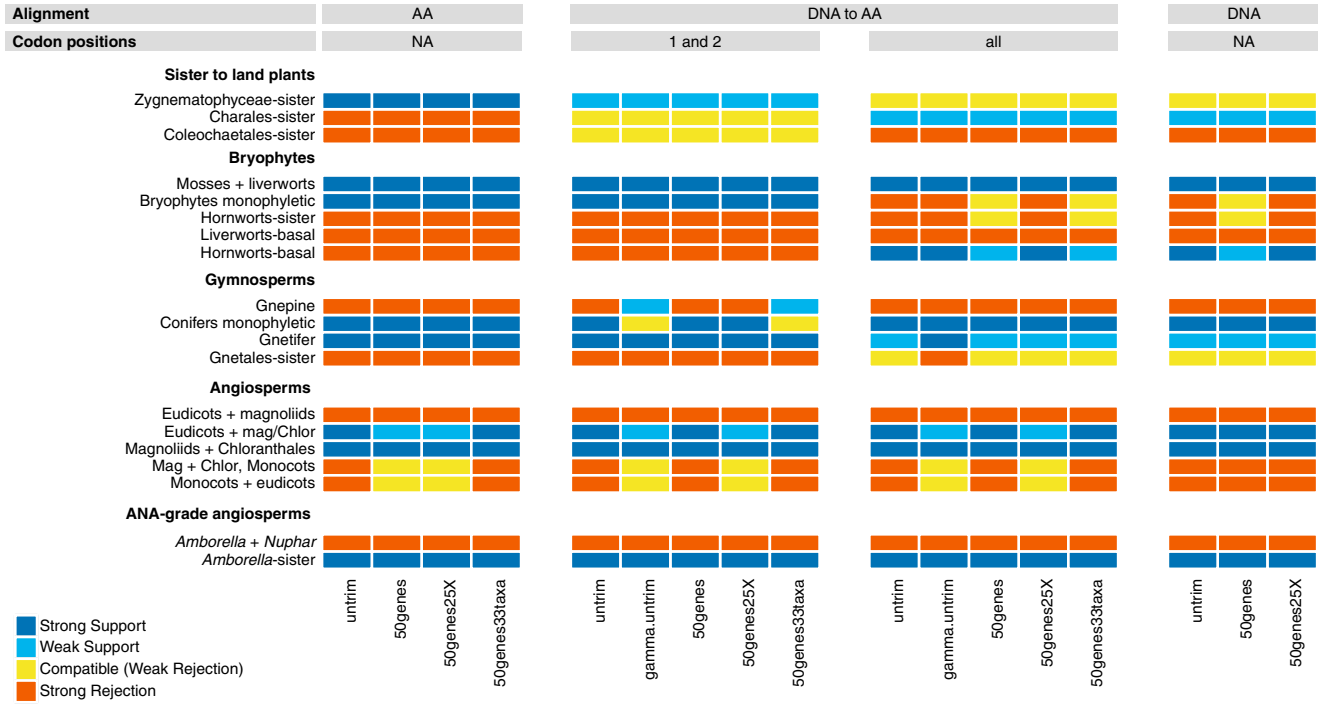

Fig. S4. Summary of support for hypotheses of land plant relationships across 17 permutations of the full data matrix are summarized for the supertree analyses. Occupancy-based gene trimming was carried out by removing genes for which >50% the full taxon set were not included in the alignment. Occupancy-based site trimming was carried out by removing sites for which >50% of the full taxon set were represented by gap characters. Long-branch trimming was performed on gene trees when a branch was 25-times longer than the median branch length. Filtering of fragmentary sequences was done by removing taxa from individual genes where they had at least 66% gaps for that gene. Filtering and analysis strategies are indicated below each column and include combinations of: (*i*) untrim: untrimmed/unfiltered data; (*ii*) 50genes: occupancy-based gene trimming at 50%; (*iii*) gamma: full Gamma (vs. PSR approximation of Gamma); (*iv*) 25X: long branch trimming; (*v*) 33taxa: filtering of fragmentary sequences. 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.

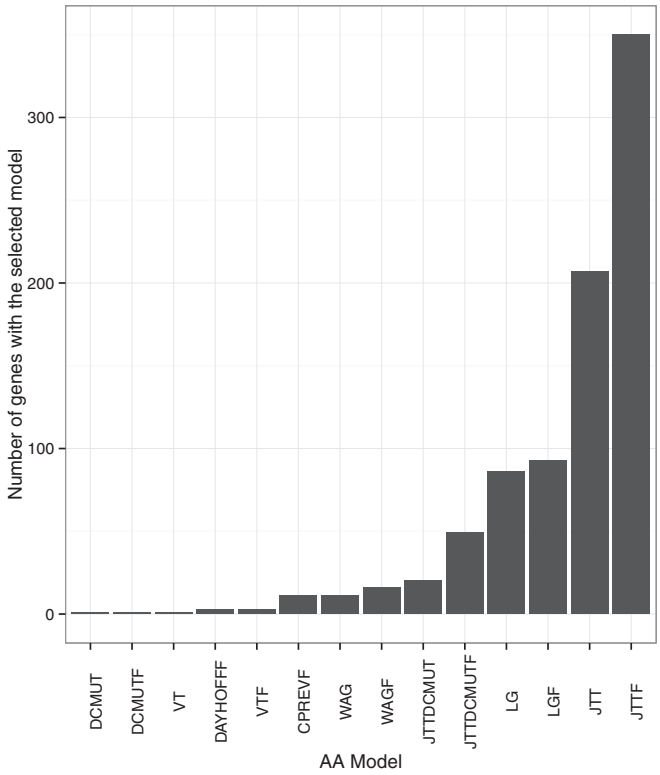

Fig. S5. Distribution of optimal amino acid substitution models across single-copy gene families. Model selection involved scoring multiple amino acid models on a fixed tree topology (Fig. 2) for each gene, and picking the model that resulted in the highest likelihood. As shown, JTTF and JTT were the highest scoring models for the majority of gene alignments.

Fig. S6. 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 number of gene trees exhibiting significant conflict (bootstrap support >75%) with each node.

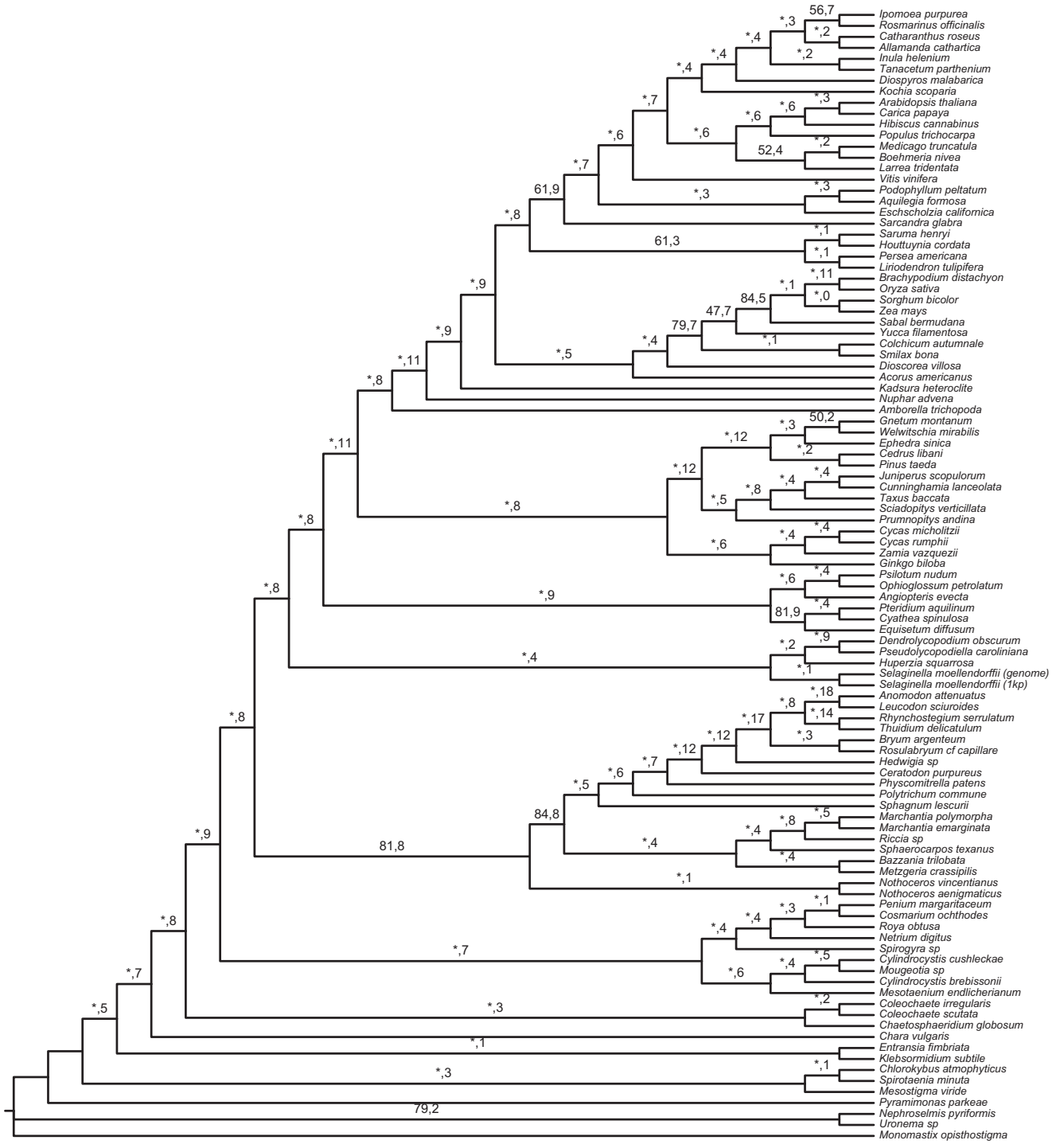

Fig. S7. ML phylogram inferred from concatenated alignments of amino acids for 674 genes after gene alignments missing more than 50% of taxa were removed and sites gapped in more than 50% of taxa were filtered. Bootstrap values for the concatenated supermatrix analysis are shown on branches (* for 100%) along with the number of gene trees exhibiting significant conflict (bootstrap support >75%) with each node.

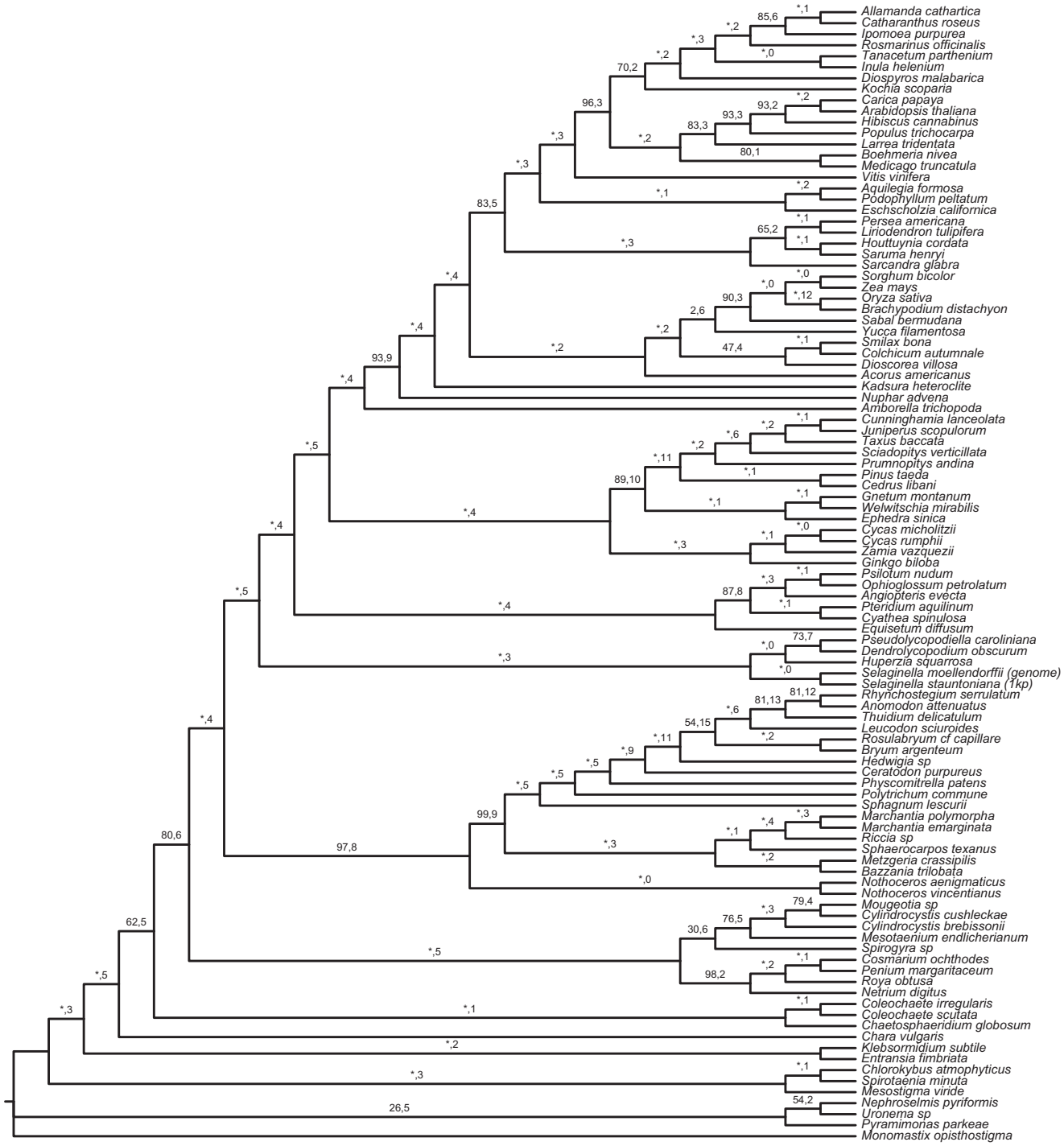

Fig. S8. ASTRAL tree estimated from 424 gene trees estimated from first and second codon position alignments after removing genes with less than 50% taxon occupancy after gene fragments missing more than 66% of their sites were removed. Bootstrap values for the ASTRAL analysis are shown on branches (* for 100%) along with the number of gene trees exhibiting significant conflict (bootstrap support >75%) with each node.

Fig. S9. ASTRAL tree estimated from 424 gene trees estimated from amino acid 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 number of gene trees exhibiting significant conflict (bootstrap support >75%) with each node.

Other Supporting Information Files

Table S1 (DOCX)
Table S2 (DOCX)
Dataset S1 (DOCX)