Syst. Biol. 63(6):862–878, 2014

DOI:10.1093/sysbio/syu049
Advance Access publication July 28, 2014

Mitochondrial Phylogenomics of Early Land Plants: Mitigating the Effects of Saturation, Compositional Heterogeneity, and Codon-Usage Bias

YANG LIU[1], CYMON J. COX[2], WEI WANG[3] AND BERNARD GOFFINET[1,*]

[1]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA;* [2]*Centro de Ciências do Mar, Universidade do Algarve, Gambelas, 8005-319 Faro, Portugal; and* [3]*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China*

**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA; E-mail: bernard.goffinet@uconn.edu.*

Received 20 January 2014; reviews returned 26 March 2014; accepted 17 July 2014
Associate Editor: Vincent Savolainen

Abstract.—Phylogenetic analyses using concatenation of genomic-scale data have been seen as the panacea for resolving the incongruences among inferences from few or single genes. However, phylogenomics may also suffer from systematic errors, due to the, perhaps cumulative, effects of saturation, among-taxa compositional (GC content) heterogeneity, or codon-usage bias plaguing the individual nucleotide loci that are concatenated. Here, we provide an example of how these factors affect the inferences of the phylogeny of early land plants based on mitochondrial genomic data. Mitochondrial sequences evolve slowly in plants and hence are thought to be suitable for resolving deep relationships. We newly assembled mitochondrial genomes from 20 bryophytes, complemented these with 40 other streptophytes (land plants plus algal outgroups), compiling a data matrix of 60 taxa and 41 mitochondrial genes. Homogeneous analyses of the concatenated nucleotide data resolve mosses as sister-group to the remaining land plants. However, the corresponding translated amino acid data support the liverwort lineage in this position. Both results receive weak to moderate support in maximum-likelihood analyses, but strong support in Bayesian inferences. Tests of alternative hypotheses using either nucleotide or amino acid data provide implicit support for their respective optimal topologies, and clearly reject the hypotheses that bryophytes are monophyletic, liverworts and mosses share a unique common ancestor, or hornworts are sister to the remaining land plants. We determined that land plant lineages differ in their nucleotide composition, and in their usage of synonymous codon variants. Composition heterogeneous Bayesian analyses employing a nonstationary model that accounts for variation in among-lineage composition, and inferences from degenerated nucleotide data that avoid the effects of synonymous substitutions that underlie codon-usage bias, again recovered liverworts being sister to the remaining land plants but without support. These analyses indicate that the inference of an early-branching moss lineage based on the nucleotide data is caused by convergent compositional biases. Accommodating among-site amino acid compositional heterogeneity (CAT-model) yields no support for the optimal resolution of liverwort as sister to the rest of land plants, suggesting that the robust inference of the liverwort position in homogeneous analyses may be due in part to compositional biases among sites. All analyses support a paraphyletic bryophytes with hornworts composing the sister-group to tracheophytes. We conclude that while genomic data may generate highly supported phylogenetic trees, these inferences may be artifacts. We suggest that phylogenomic analyses should assess the possible impact of potential biases through comparisons of protein-coding gene data and their amino acid translations by evaluating the impact of substitutional saturation, synonymous substitutions, and compositional biases through data deletion strategies and by analyzing the data using heterogeneous composition models. We caution against relying on any one presentation of the data (nucleotide or amino acid) or any one type of analysis even when analyzing large-scale data sets, no matter how well-supported, without fully exploring the effects of substitution models. [Compositional heterogeneity; early land plants; evolutionary saturation; mitochondrial genome; phylogenomics; synonymous codon-usage bias.]

Phylogenomic data are typically considered as the panacea for resolving deep phylogenetic ambiguities based on the assumption that the accurate phylogenetic signal would emerge from large quantities of data (Rokas et al. 2003; Dunn et al. 2008). Phylogenomic data are, however, typically just a concatenation of a suite of discrete loci, each characterized by evolutionary constraints that shape its substitution pattern, and may result in homoplasious substitutions that may mislead phylogenetic inferences. Consequently, these types of data should not be considered free of phylogenetic noise and the mere size of the data sets is not sufficient to guarantee accurate reconstruction (Philippe et al. 2011) as the phylogenetic effects of homoplasy within individual loci that accumulate in genomic data may strongly bias inferences (Delsuc et al. 2005). Processes causing homoplasy in nucleotide (nt) gene data may include substitutional saturation (repeated substitutions at a single site, a factor of time and selective functional constraints, see Jeffroy et al. 2006; Philippe et al. 2011), among-lineage compositional heterogeneity (i.e., GC content heterogeneity, see Foster 2004; Inagaki et al. 2004; Inagaki and Roger 2006; Regier et al. 2010; Rota-Stabelli et al. 2013), or codon-usage bias (distinct preferences for alternative synonymous codons, see Gouy and Gautier 1982; Stenøien 2005).

Substitutional saturation is the most frequently discussed cause of phylogenetic artifacts. If multiple substitutions occur frequently at the same position of an alignment, the apparent pairwise genetic distances underestimate the real distances and the alignment is said to be saturated (Philippe et al. 2011). In a saturated data matrix, synapomorphies may be erased by additional mutations, which can weaken branch support, and potentially mislead phylogenetic inferences (Jeffroy et al. 2006). Although mitochondrial (mt) genes evolve slowly in plants (Palmer and Herbon 1988), they are not free of saturation, especially when

comparing anciently diverged sequences. Nucleotide and protein sequences may also exhibit among-taxa compositional heterogeneity (i.e., nonstationarity). Stationary substitution models assume composition homogeneity and may cluster unrelated taxa sharing similar compositions, thus lead to phylogenetic artifacts (Lockhart et al. 1992; Mooers and Holmes 2000; Blanquart and Lartillot 2008). The composition of nucleotide sequences is correlated to a preference in usage of synonymous codons (Stenøien 2005). The lineage-specific codon-usage biases, which have been observed in many organisms (Liu et al. 2004; Zhou and Li 2009; Wang et al. 2010; Plotkin and Kudla 2011), may have strong effects on phylogenetic inferences (Foster 2004; Inagaki et al. 2004; Inagaki and Roger 2006; Regier et al. 2010). Depending on the extent of convergence in codon usage, taxa that are unrelated but share similar patterns of codon preference may be resolved as having arisen from a unique common ancestor (Rota-Stabelli et al. 2013).

Land plants (embryophytes) are the primary producers of terrestrial ecosystems. Their conquest of land can be traced back to late Silurian (428–423 Ma) (Kenrick et al. 2012). The fossil record is particularly challenging to interpret, since remains in older sediments are microscopic and their affinities to land plants are somewhat ambiguous (Edwards et al. 2014). The early diversification of land plants may have been rapid (Kenrick and Crane 1997; Lewis et al. 1997; Bateman et al. 1998; Kenrick et al. 2012), and the reconstruction of such ancient cladogenic events that occurred over short time frames is difficult to reconstruct as their signatures may be scarce and erased by subsequent substitutions (Whitfield and Lockhart 2007). Despite numerous attempts to reconstruct the deep relationships among land plants, the order of their earliest divergences remains ambiguous (Goffinet 2000; Qiu 2008). Among the extant embryophytes, the ambiguity rests primarily in the relationships of bryophyte lineages, that is, of liverworts, mosses, and hornworts. Whether the three lineages arose from a unique common ancestor (e.g., Garbary et al. 1993; Renzaglia et al. 2000), or compose a grade leading to vascular plants, or more accurately Polysporangiophytes (e.g., Kenrick and Crane 1997), and if so, in what order (e.g., Goffinet 2000) is still unresolved. Within a paraphyletic scenario of bryophytes giving rise to vascular plants, either liverworts (Mishler and Churchill [1984], morphological data; Bremer [1985], morphological data; Kenrick and Crane [1997], Lewis et al. [1997], chloroplast [cp] *rbc*L gene; Qiu et al. [1998], mt genome structure; and Kelch et al. [2004], cp genome structure) or hornworts (Malek et al. [1996], mt *cox*3 gene; Garbary and Renzaglia [1998], morphological data; and Duff and Nickrent [1999], mt SSU gene) have been resolved as sister to the remaining land plants. Inferences from single or combined loci resolved liverworts and mosses as sharing a unique ancestor, with hornworts either sister to all remaining land plants or only to vascular plants (Hedderson et al. 1996; Nishiyama and Kato 1999; Nickrent et al. 2000; Renzaglia et al. 2000). A monophyly of bryophytes has been supported by ultrastructural (Garbary et al. 1993) and 5S rRNA sequences (Hori et al. 1985) and perhaps also by theoretical considerations of the origin of alternate life cycles from an ancestor with isomorphic generations (Taylor et al. 2005; Goffinet and Buck 2013).

The evolution of land plants and their sister-groups has been reconstructed recently from variation in genomic-scale data sampled from the chloroplast (Cox et al. 2014; Ruhfel et al. 2014; Zhong et al. 2014), mitochondrial (Turmel et al. 2013), or nuclear (Finet et al. 2010; Wodniok et al. 2011; Zhong et al. 2013) genomes. These phylogenomic inferences were, however, drawn from a limited bryophyte taxon sampling (i.e., often with only one exemplar per lineage and sometimes without all three bryophyte lineages included). The currently, most widely accepted hypothesis whereby liverworts are sister to the remaining land plants, and mosses diverged from the common ancestor shared by hornworts and extant vascular plants, is the one inferred by Qiu et al. (2006a; 2007) from discrete loci drawn from all three genomic compartments sampled for an extensive set of taxa. However, despite the extensive taxon sampling and the concatenation of multiple loci, topological ambiguities persist as reflected in the incongruence between plastid nucleotide data and their amino acid translations (Qiu et al. 2006b; Cox et al. 2014).

The recent advances in high-throughput sequencing techniques that mark the beginning of the phylogenomic era (Delsuc et al. 2005) provide an opportunity to reconstruct, based on extensive data sets, ancient events that are critical to our understanding of the evolution and diversification of life, such as the origin and radiation of early land plants. Plant mt genomes appear well suited for deep phylogenomic inferences: they comprise about 40 genes and are now easily sequenced (e.g., Zhang et al. 2011; Liu et al. 2013), and their rate of substitution is much lower than that of chloroplast (<3×) or nuclear (<10×) genes (Wolfe et al. 1987; Palmer and Herbon 1988; Graur and Li 2000; Drouin et al. 2008), which may result in less saturation, and hence less homoplasy (Nickrent et al. 2000; Qiu et al. 2010). Here, we present the first inference of the deep relationships among major lineages of extant land plants using the entire sets of 41-mt protein-coding genes sampled for 60 streptophytes, including 20—newly sequenced—bryophytes. We provide further evidence that caution must be used when relying solely on inferences drawn from large amount of protein-coding nt data, as substitutional saturation, among-lineage compositional biases, and codon-usage biases may lead to highly supported yet erroneous hypotheses.

Materials and Methods

Taxon Sampling

Mitochondrial protein-coding genes from 20 bryophytes, including three liverworts representing

two orders, and 17 mosses representing 11 orders, were newly sequenced using the next-generation sequencing method (Supplementary Table S1, http://datadryad.org/resource/doi:10.5061/dryad.7b470). Combined with the bryophytes for which the mt genomes were previously sequenced, the taxon sampling covers 5 of the 15 recognized liverwort orders (Crandall-Stotler et al. 2009), 13 of the 29 moss orders (Goffinet et al. 2009), and 2 of the 5 hornwort orders (Renzaglia et al. 2009). In addition, the mt genomes of various streptophytes (i.e., 5 algae, 2 lycophytes, 1 gymnosperm, and 25 angiosperms) were downloaded from the NCBI's genome database (http://www.ncbi.nlm.nih.gov). In all phylogenetic analyses, we used *Mesostigma viride* and *Chlorokybus atmophyticus* as outgroups, based on (Karol et al., 2010). The matrix therefore comprised 60 streptophytes, including 27 nonvascular land plants, 28 vascular plants, and 5 algae. We identified 41 protein-coding genes from the mt genomes of these taxa. Genes that were pseudogenized or lost were coded as missing data (Supplementary Table S2).

DNA Extraction, Next-Generation Sequencing, and Mitochondrial Data Assembly

Approximately 0.5–3.5 g dried or fresh gametophyte and/or sporophyte were sampled from each recently collected population of bryophyte. The total cellular DNA was extracted following the steps described by Liu et al. (2013). DNA quantity and quality were assessed using the Qubit fluorometer system (Invitrogen, San Diego, CA, USA) and NanoDrop Spectrophotometer ND-1000 (NanoDrop Technologies, Wilmington, DE, USA). Genomic DNAs from different taxa were multiplexed and sequenced either on the GS FLX 454 (Roche 454 Life Science, Branford, CT, USA) at the IGSP Sequencing Core Facility (Duke University) or on Illumina HiSeq 2000 Sequencing System (Illumina, CA, USA) at the Health Center Translational Genomics Core Facility (University of Connecticut). *De novo* assembly and reference-guide mapping of the raw reads were performed using the software CLC Genomics Workbench (CLC Bio, Aarhus, Denmark). Mitochondrial contigs were aligned against the published mt genomes of liverworts (*Marchantia polymorpha*, Oda et al. [1992]; *Pleurozia purpurea*, Wang et al. [2009]; *Treubia lacunosa*, Liu et al. [2011]) or mosses (*Physcomitrella patens*, Terasawa et al. [2007]; *Anomodon rugelii*, Liu et al. [2011]) using the genome alignment software Mauve v2.3.1 (Darling et al. 2004) nested inside the program Geneious v6.0.3 (http://www.geneious.com/; Biomatters Ltd, Auckland, New Zealand). Draft mt genomes were annotated in Geneious. Protein-coding sequences were extracted by custom PERL scripts (Supplementary Material).

Sequence Alignment and Gene Evolutionary Rate Analyses

Mitochondrial protein-coding genes were aligned individually with TranslatorX (Abascal et al. 2010) using MAFFT (Katoh et al. 2005) to compute the amino acid (aa) alignments. In all cases, poorly aligned amino acid regions were trimmed from alignment by GBLOCKS with the least stringent settings (Talavera and Castresana 2007), so that nucleotide (nt) alignments are produced based on corresponding amino acid alignments after removal of ambiguous positions. The above processes were automatically conducted on TranslatorX web server (http://translatorx.co.uk/). After removing stop codons, 41 single-gene nucleotide alignments were concatenated into a master alignment using Geneious v6.0.3, and converted into appropriate formats using PAUP* v4.0b10 (Swofford 2003). Thereafter, the nucleotide data were directly translated to amino acids based on the standard genetic code in MEGA v5 (Tamura et al. 2011). Both the nucleotide and amino acid data were used for subsequent phylogenetic analyses. All newly generated mt protein-coding sequences have been deposited in the GenBank. Voucher information and GenBank accession numbers are provided in Supplementary Table S1. Mitochondrial genes may possess RNA editings, and the number of editing sites may be considerable in plants (Knoop 2004), although few are conserved across lineages (Finster et al. 2012). Large-scale analyses in seed plants (Qiu et al. 2006a) or angiosperms (Qiu et al. 2010) indicated that RNA editings had no obvious effect on phylogenetic reconstructions. Previous studies showed that gene rate heterogeneity may play an important role in phylogenetic reconstructions, and splitting the data matrix by rate categories may result in different topologies (Jian et al. 2008; Barrett et al. 2012). To assess the rate heterogeneity of mt genes, we calculated each gene's average pairwise (p) distance in MEGA 5 (Tamura et al. 2011).

Single-Gene Tree Inferences and Supernetwork

Single-gene phylogenies often produce incongruences in phylogenomic analyses (Jeffroy et al. 2006), which may be due to (i) stochastic errors owing to insufficient data or (ii) violations of the orthology assumption generated by mechanisms such as incomplete lineage sorting (Whitfield and Lockhart 2007). However, given that mt genomes are uniparentally inherited, no lineage sorting problem is expected. To evaluate the consistency among single genes, maximum-likelihood (ML) analyses were performed for each gene alignment using RAxML v7.2.3 (Stamatakis 2006) under the GTR + I + Γ model and bootstrap support was estimated based on 100 pseudoreplicates using a GTR–CAT approximation. To visualize conflicts among genes, we constructed a supernetwork (allowing for missing taxa) for the 41-mt gene trees using SplitsTree v4.11.3 (Huson et al. 2004; Huson and Bryant 2006). As single-gene trees usually lack statistically significant support, to reduce the risk of overestimating the conflicts among single gene trees, only nodes with significant support (BS $\geq$ 85%) were preserved and all other nodes were collapsed.

An 85% majority-rule consensus of the 100 bootstrap trees from each gene analysis was generated in PAUP* v4.0b10 (Swofford 2003), and the 41 consensus trees used as the input data for SplitsTree. When calculating the supernetwork, we used the Z-closure option, mean edge weights, set splits transformation as equal angle, and left all other parameters as default settings.

Phylogenetic Analyses Based on Standard Homogeneous Models

Both the concatenated nucleotide and the corresponding translated amino acid data sets were analyzed with ML and Bayesian inference (BI) under a homogeneous model, that is, assuming composition homogeneity among taxa. For the nt data set, ML analyses were performed using the parallel version of RAxML v7.2.3 (Stamatakis 2006). The ML trees were calculated under the GTR + I + Γ model. Nonparametric bootstrap analyses were implemented by GTR–CAT approximation for 100 pseudoreplicates. Former studies showed data partitioning strategies may affect the topology and nodal supports estimates (Brown and Lemmon 2007; McGuire et al. 2007; Li et al. 2008). To determine the optimal partitioning of the data, we explored six distinct partition schemes in ML analyses. The data set was partitioned *a priori* on the basis of gene identity and evolutionary constraints (e.g., codon positions). The six partitioning schemes are: single (not partitioning); 12-3 (first and second codon positions combined in one partition; third codon positions as one partition); 1-2-3 (each first, second, and third codon position as one partition); gene (partitioned by gene); gene-12-3 (first and second codon positions in each gene as one partition, and third in each gene as one partition); gene-1-2-3 (each of the three codon positions in each gene as one partition). The optimal partitioning strategy was determined by Akaike Information Criterion (AIC) (Akaike 1974) and Bayesian Information Criterion (BIC) (Schwarz 1978). Unpartitioned homogeneous BIs using a GTR + I + Γ substitution model were conducted using MrBayes v3.2 (Ronquist et al. 2012) and P4 v0.90 (Foster 2004). In addition, MrBayes analysis was run with the data partitioned by codon position (3× [GTR + I + Γ]). Markov Chain Monte Carlo (MCMC) was run for 2–5 million generations in each analysis. The model in all P4 analyses included a "polytomy prior," which describes a proposal to allow polytomies within the tree using a "resolution class" (Lewis et al. 2005) with a strong prior probability favoring polytomies (C = log(10)). For partitioned analyses, substitution model parameters were unlinked among the partitions, so that they were estimated independently for each partition. In all analyses, branch lengths and topology were linked. Burn-in and convergence were assessed using the likelihood of the samples plotted against generation time and by monitoring diagnostics within and between chains from the estimated posterior distribution (for a detailed description, see Supplementary Information). Posterior probabilities (PPs) of clade support were estimated by sampling trees from the posterior distribution after removal of the burn-in samples. All estimates of marginal likelihoods were computed in P4, which implements equation (16) of (Newton and Raftery, 1994). This estimator of the marginal likelihood was formulated to overcome some of the problems associated with the harmonic-mean estimator (Kass and Raftery 1995) but is still known to be inaccurate in some circumstances (Lartillot and Philippe 2006).

Due to the high dimensionality of amino acid substitution models, empirically estimated models are typically used to reduce the computational burden. However, although using a model that fits the data well is essential, existing mitochondrial models, such as MtREV (Adachi and Hasegawa 1996), are all based on data from animal mt genomes. We therefore estimated a new amino acid substitution model for the current plant data using the ML approach implemented in the ATGC bioinformatics platform (Dang et al. 2011). The empirical exchange rates and composition frequencies were calculated from the entire streptophyte amino acid alignment after excluding ambiguous regions. The resulting protein model is here named stmtREV, as it is derived from streptophyte mitochondrial data and is a general time-reversible model (stmtREV_model.txt, Supplementary Material). Additionally, the amino acid data set was run through the program modelgenerator (Keane et al. 2006), which measures the goodness of fit of the data to existing empirical models. The best-fitting standard empirical model, JTT (Jones et al. 1992), and the newly estimated stmtREV were used in Bayesian analyses, and both models, plus the GTR model, were applied in ML analysis of the amino acid data.

Phylogenetic Analyses Using Heterogeneous Composition Models

The presence of among-lineage compositional heterogeneity may compromise the accuracy of phylogenetic inferences when using standard homogeneous substitution models resulting in taxa with similar compositions being artificially attracted to each other (Lockhart et al. 1994; Yang and Roberts 1995; Foster 2004). To test for the presence of such artifacts in our analyses, we performed nonstationary composition MCMC in P4 on both the nucleotide and amino acid data using the node-discrete composition heterogeneity (NDCH) model (Cox et al. 2008), which allows the composition to vary among lineages (Foster 2004). Additional composition vectors were added to the substitution models and the fit of the model to the data assessed using posterior predictive distribution tests of the χ^2 statistic of composition homogeneity (Foster 2004). All P4 analyses included a polytomy prior as described previously. NDCH model analyses of nt data were performed using the GTR + I + Γ substitution model and analyses of the amino acid data

were performed using the JTT + Γ and stmtREV + Γ substitution models.

Similarly, among-site composition heterogeneity can cause phylogenetic artifacts (Lartillot and Philippe 2004). To test for among-site compositional heterogeneity in both the nt and amino acid data, we performed MCMC analyses using the CAT + GTR + Γ model implemented in PhyloBayes MPI v1.2d (Lartillot et al. 2009).

Evaluation of Substitutional Saturation, Compositional Heterogeneity, and Codon-Usage Bias within Nucleotide Data Set

To estimate the amount of substitutional saturation, we plotted the uncorrected *P*-distances against the inferred distances using the method described by Philippe and Forterre (1999). The level of saturation is estimated by computing the slope of the regression line in the plot, the shallower the slope, the greater the degree of saturation. We estimated saturation for two subsets of the concatenated data: the combined first and second codon positions, and the third codon positions. Furthermore, to evaluate the saturation level in the backbone phylogeny, we isolated pairs of taxa representing two major lineages. To determine the effect of saturation on the phylogenetic reconstructions, all third codon positions, and the first and second codon positions, were analyzed independently. ML analyses were performed for the two subsets, in each case with the data partitioned by genes.

Base compositions of each gene and each codon position were obtained from the output of the alignment generated by the software TranslatorX. To visualize compositional heterogeneity, the GC percentage score of each species was plotted. To characterize the differences in composition heterogeneity among groups, we performed pairwise one-tailed student *T*-tests between any two major plant lineages using the R package (http://www.r-project.org/). Although the data are linked through the phylogeny and thus not independent, this test offers a conservative estimate of the significance of the differences of the among-group compositional heterogeneity.

To evaluate synonymous codon bias—whether synonymous codons are used in different amounts—among land plant lineages, we performed a correspondence analysis of Relative Synonymous Codon Usage (RSCU) values as implemented in the software GCUA (McInerney 1998). To evaluate the impact of the nucleotide substitution patterns underlying the observed codon-usage biases, the nucleotide data were "degenerated," whereby synonymous codons were reduced to a single triplet, composed of ambiguous sites reflecting codon redundancy (Criscuolo and Gribaldo 2010; Regier et al. 2010). For example, the synonymous codons AAA and AAG encode Lysine, and all occurrences of these two codons in the data matrix were recoded as AAR. Similarly, six synonymous codons for Serine (i.e., TCT, TCC, TCA, TCG, AGT, and AGC) were replaced with WSN in the data matrix (for all re-coding, see Cox et al. 2014). Both ML and BI analyses were performed for this degenerate nucleotide (deg-nt) data. In the ML analyses, six partition schemes (as described above for the original nt data) were also evaluated using the GTR + I + Γ substitution model. Homogeneous Bayesian MCMC were performed using MrBayes and P4 with a GTR + I + Γ model.

Alternative Hypothesis Testing

To test the various hypotheses proposed by former studies, that is, (i) liverworts, (ii) mosses, or (iii) hornworts are sister to the remaining land plants, (iv) liverworts and mosses share a unique common ancestor, and (v) bryophytes (including liverworts, mosses, and hornworts) are monophyletic, we performed both nonparametric and parametric bootstrapping tests using nt and aa data. In the nonparametric bootstrapping test, the significances of departures of alternative hypotheses were assessed by the Kishino–Hasegawa (KH) test (Kishino and Hasegawa 1989), the Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa 1999), and the approximately unbiased (AU) test (Shimodaira 2002). In each case, the constrained tree (with other lineages composing a polytomy) was constructed in Mesquite v2.74 (Maddison and Maddison 2001), and then optimized in RAxML under GTR + I + Γ model (nt data) or stmtREV (aa data) model. All the tests were conducted using TREE-PUZZLE v5.2 (Schmidt et al. 2002) and CONSEL v0.1k (Shimodaira and Hasegawa 2001). Site-wise log likelihoods were estimated by TREE-PUZZLE, and the values were also used as input data for CONSEL. For these nonparametric bootstrapping tests, only the five hypothetical macroevolutionary topologies mentioned above were compared.

Parametric bootstrapping analyses were performed using the Swofford–Olsen–Waddell–Hillis (SOWH) test (Goldman et al. 2000). The SOWH test is considered more powerful than the nonparametric bootstrap tests as it can avoid Type I error when model parameters are accurately provided (Buckley 2002). Data are simulated based on the alternative hypothesis constrained tree. For nt data, we constrained the tree as for the nonparametric bootstrapping analyses, and then optimized the tree topology, branch lengths, and model scores from the original data in RAxML, under GTR + I + Γ model and the best partitioning scheme. We simulated 100 replicates of nucleotide data sets based on this optimized tree using SEQ-GEN v1.3.2 (Rambaut and Grass 1997). Each partition was simulated individually based on its own parameters and alignment length. Finally, all partitions in a certain replicate were concatenated into one data set. For amino acid data, the constrained trees were first optimized in RAxML with stmtREV model, and then 100 amino acid data sets were simulated using the program Indelible v1.03 (Fletcher and Yang 2009). All ML analyses in SOWH test were performed under homogeneous models only. The optimized tree (with branch lengths)

and stmtREV substitution matrix were used as input data. For each of these simulated data sets, we conducted two likelihood searches, one to find the optimal unconstrained tree and the other to find the optimal constrained tree. The distribution of log-likelihood differences between each "optimal constrained" and "optimal unconstrained" trees was generated and used for evaluating the significance of the difference between the best tree and constrained tree. To perform the calculations, custom scripts were created to automate the procedures of extracting the data matrix from SEQ-GEN output files (extractdata.py; Supplementary Material), and extracting likelihood scores from RAxML results (extract likelihood; Supplementary Material). The distributions of likelihood differences were plotted in Excel. Ultimately, the difference between the likelihood scores of the unconstrained and constrained trees inferred using the original data set was compared with the distribution of likelihood differences between trees inferred from the simulated data sets under *P*-value of 0.05.

To assess the potential influence of taxon sampling, four-cluster likelihood mapping analyses were performed using PUZZLE 4.0 (Strimmer and von Haeseler 1997). In the analyses, four groups of taxa were defined, based on the four major land plant lineages, namely liverworts, mosses, hornworts, and vascular plants. Each time, four taxa were sampled from each of the four groups, and relative likelihood frequency for each of the three possible topologies was calculated. Likelihood frequencies were mapped on a triangle partitioned into three regions, each region corresponding to one of the three possible topologies. The analyses were performed under the GTR + Γ model with eight discrete gamma categories, and sampled for 20,000 randomly quartets.

Results

Data Sets

Our data matrix consists of 60 taxa and 41-mt protein-coding genes (Supplementary Table S1). After the removal of ambiguous positions, the concatenated nucleotide data set comprised 31,467 characters. As a direct translation of the nucleotide alignment, the amino acid data set includes 10,489 characters. Except for hornworts, which lost half of their mt protein-coding genes and are thus represented by 20 genes only, most land plants have more than two-thirds of the 41 genes (Supplementary Table S1). Because the ribosomal protein genes (*rps* and *rpl*) tend to be lost in many plants, these genes are present in the fewest exemplars. Except for *mat*R, all genes are present in at least one taxon from two or more of the four major lineages of land plants. In total, 16% sites in the matrix are represented by missing data, representing primarily gene losses (Supplementary Table S2).

The average *P*-distances of the 41 individual mt gene alignments ranged from 0.08% to 0.27%, *atp*8 exhibits most variation, and *mat*R the least, but it should be noted that *mat*R is only present in vascular plants. Based on the average *P*-distances, no clear gene rate category could be recognized (Supplementary Fig. S1). Thus, our data set was not partitioned or split based on the rate category of genes in the subsequent analyses.

Single-Gene Trees: Incongruences and Supernetwork

Single-gene tree inferences yielded only weak to moderate support values for nodes defining relationships among the major lineages of land plants (Supplementary Fig. S2). The supernetwork analysis indicates that each major group of land plants, except for lycophytes, is always resolved as monophyletic (Fig. 1). Single-gene trees differ in the relationships among or within major groups, but never in the circumscription of the four major lineages (i.e., liverworts, mosses, hornworts, and vascular plants), suggesting that none of the genes have been transferred horizontally between these lineages.

Phylogenetic Results under Homogeneous Models

For the nt data, six partitioning schemes were tested by ML analyses. Partitioning has no obvious effect on tree inferences, as all partitioning schemes generated congruent topologies with similar nodal supports (data not shown). The partitioning strategy gene-12-3 is favored by both AIC and BIC in the nt data (Supplementary Table S3).

Homogeneous analyses of either nt or aa data yield maximal support ($ML_{nt\&aa}$-BS = 100%, $BI_{nt\&aa}$-PP = 1.0; Fig. 2) for the monophyly of land plants, and its major four groups (i.e., liverworts, mosses, hornworts, and vascular plants). The relationships among these groups differ, however, based on the character source: mosses are resolved as sister to the remaining land plants when inferences are drawn from nucleotide data (ML-BS = 70%, BI-PP = 1.0; Fig. 2a), whereas liverworts occupy that position based on the amino acid data (ML-BS = 82%, BI-PP = 1.0; Fig. 2b). Except for this conflict, nt and aa trees are congruent, in terms of estimating the detailed topologies within each group. Both nt and aa data resolved hornworts as sister to vascular plants (ML_{nt}-BS = 82%, BI_{nt}-PP = 1.0; ML_{aa}-BS = 98%, BI_{aa}-PP = 1.0; Fig. 2a,b), which is consistent with the four-cluster likelihood mapping analyses (nt: 95.9% and aa: 97.5%, Fig. 2c,d). Bayesian analyses using alternative models for both nt (Supplementary Fig. S9 codon site-specific model [3× (GTR + I + Γ)]) and aa data (Supplementary Fig. S10 stmtREV + Γ + PP model) did not resolve the conflict as each data type supported similar trees to the models used in the analyses presented in Figure 2.

Phylogenetic Results under Heterogeneous Models

Bayesian analyses of nt data using P4 under nonstationary composition model estimated liverworts

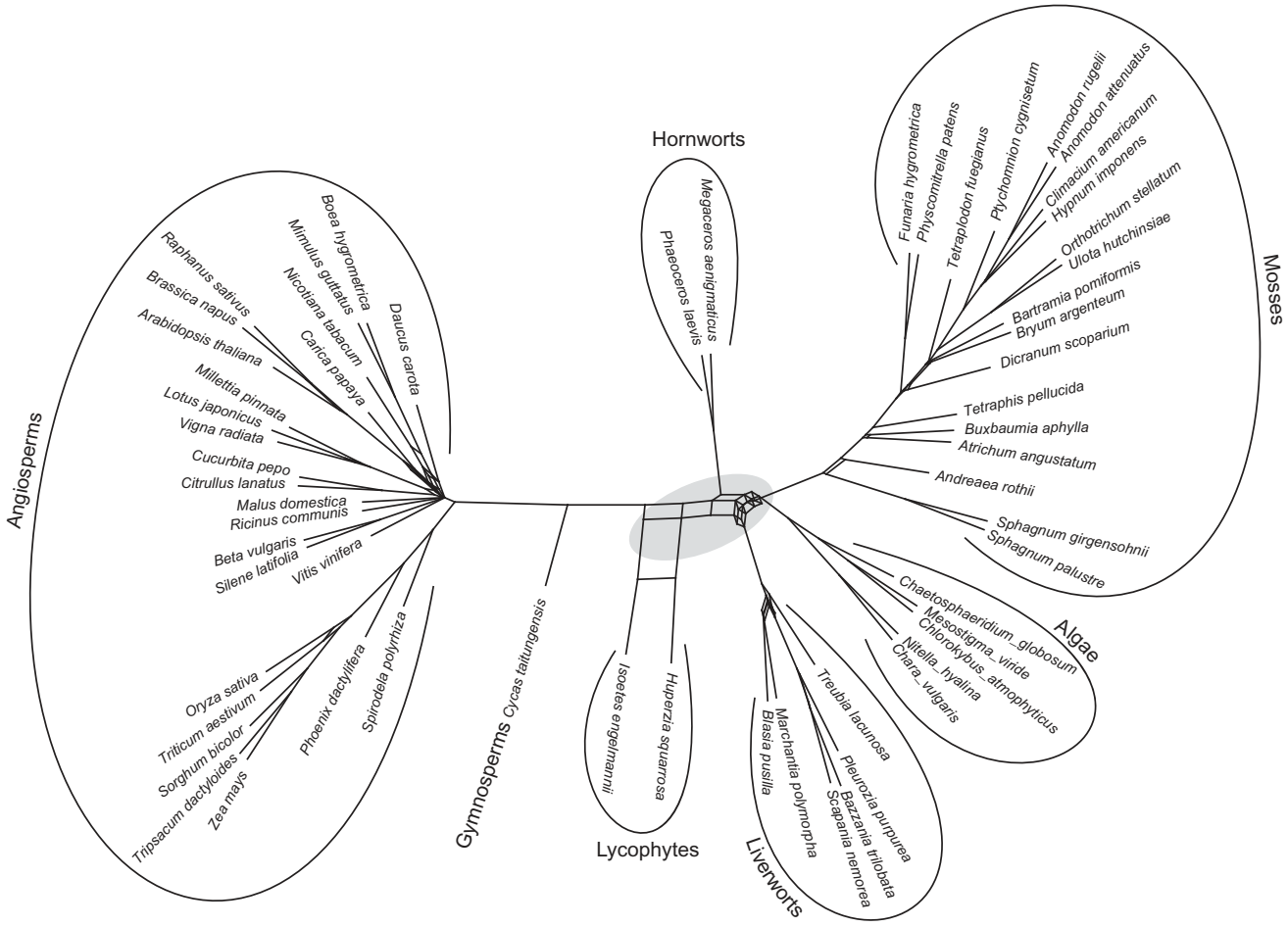

FIGURE 1. A supernetwork of 41 land plant mitochondrial single-gene trees. The network was constructed using SplitsTree. The 85% majority-rule consensus of 100 bootstrap trees was used as the input tree for each gene, which means only strong conflicts (BS ≥ 85%) are presented on the network. Parallelograms indicate incongruence among single-gene trees. Gray-shaded area encompasses the competing splits among major streptophyte lineages.

as the earliest land plant lineage but with low probability (NDCH-BI_{nt}-PP = 0.70; Fig. 3 and Supplementary Fig. S11), indicating that support for the earliest branching of mosses using homogeneous models is in part due to among-lineage compositional bias. P4 analyses of aa data using nonstationary composition model also supported liverworts as sister to the remaining land plants with high probability (NDCH-BI_{aa}-PP = 1.0; Supplementary Fig. S12), which is the same result as that obtained by analyses of the aa data under the homogeneous models (Supplementary Figs. S6 and S10). Under a nonstationary composition model, nt and aa data are thus congruent, with aa alone strongly supporting an early-branching liverworts.

By contrast, Bayesian analyses of nt data using PhyloBayes strongly supported an early-branching of the mosses (CAT-BI_{nt}-PP = 1.0; Supplementary Fig. S13) as did homogeneous model analyses of the same data (Supplementary Figs. S5 and S7). PhyloBayes analyses of the amino acids identified an early-branching liverworts although at a low probability (CAT-BI_{aa}-PP = 0.66; Supplementary Fig. S14), indicating that support for an early-branching liverworts in site-composition homogeneous analyses (Supplementary Figs. S8 and S10) and nonstationary composition analyses (Supplementary Fig. S12) of the amino acid data may in part be due to among-site compositional biases. Under a composition site-heterogeneous model, nt and aa data are thus congruent, with nt alone strongly supporting an early-branching mosses.

Saturation, Compositional Heterogeneity, and Codon-Usage Biases within nt Data Set

The degree of substitutional saturation was estimated at the third and combined first + second codon positions. Third codon positions (slope = 0.35; Fig. 4b) are more substitutionally saturated than first + second codon positions (slope = 0.75; Fig. 4a). The degree of saturation is even more pronounced if calculated based on comparisons of any two taxa representing two of the four major lineages of land plants (i.e., slope = 0.25 vs. 0.35 in third codon, Fig. 4d vs. Fig. 4b and 0.67 vs. 0.75 in first + second codon, Fig. 4c vs. Fig. 4a).

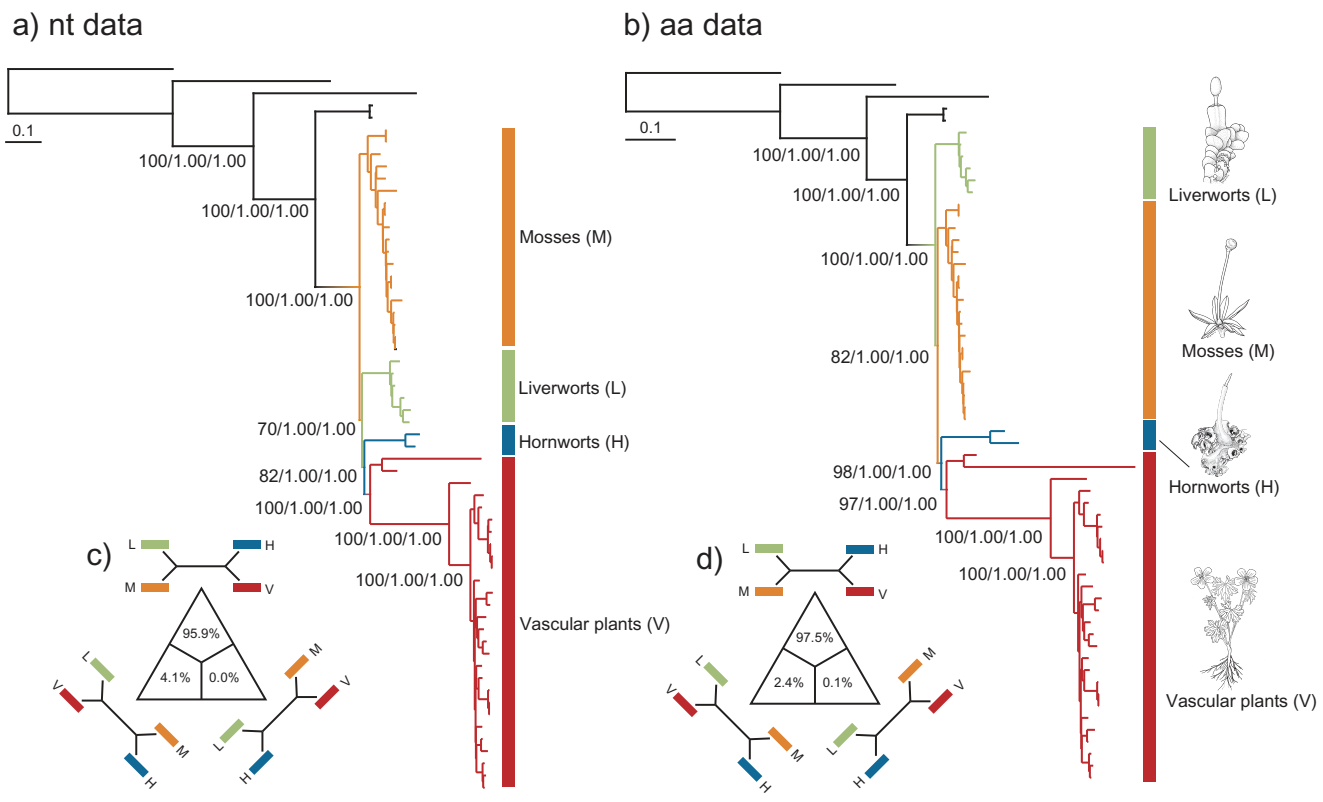

FIGURE 2. Maximum-likelihood phylograms of land plants inferred from concatenation of 41 mitochondrial genes under a homogeneous model: a) nt data and b) aa data. Nodal supports for the backbone relationships are exhibited. The support values are summarized from the bootstrap-likelihood analyses (nt, Supplementary Fig. S3 and aa, Supplementary Fig. S4), and PPs from BIs using MrBayes (nt [GTR + I + Γ], Supplementary Fig. S5; aa [JTT + Γ], Supplementary Fig. S6) and P4 (nt [GTR + I + Γ + PP], Supplementary Fig. S7; aa [JTT + Γ + PP], Supplementary Fig. S8), respectively. The trees exhibit incongruence between the nucleotide and amino acid data set by mosses or liverworts being sister to the rest of land plants. c) and d) are results from the four-cluster likelihood mapping analyses from the respective nt and aa data. In each case, the sequences were split into four groups corresponding to the major land plant lineages. The corners of the triangles are labeled with the three alternative topologies.

Inferring the relationships based on ML of only third codon positions yields mosses as the sister-group to the remaining land plants with weak support (ML_{3rd}-BS = 64%; Supplementary Fig. S15), whereas the signal extracted from the first + second codon positions favored liverworts in that position also with a low probability ($ML_{1st+2nd}$-BS = 54%; Supplementary Fig. S16).

Vascular plants and nonvascular plants significantly differ in their GC composition across the entire coding region or individual codon positions (Fig. 5 and Supplementary Table S4). Vascular plants have the highest overall GC content (42.70%) and are on average 27% higher than that of algae, which have the lowest GC percentage. Within nonvascular plants, liverworts have the highest overall and third codon GC content, and mosses have the lowest, the differences between them are significant based on the statistical student *T*-test (*P*-value ≤ 0.05, Supplementary Table S4). Certain liverworts exhibit high levels of GC in the third codon positions, similar to those observed for vascular plants (Fig. 5). Hornworts are not significantly different from liverworts, mosses, or vascular plants when the overall, first or second codon positions are evaluated, and differ only significantly in the third codon position when compared with overall vascular plants. The small sample size of hornworts (only two taxa) may make the *T*-test inaccurate. In all organisms, the first codon positions have the highest and the third codon positions the lowest GC content.

RSCU values reveal distinct patterns of codon-usage among algae, bryophytes, and tracheophytes. Algae and tracheophytes exhibit the most divergent usage, and bryophytes show intermediate preferences with codon usage in mosses, rather than liverworts being slightly more similar to that of algae (Fig. 6). Inferences from degenerated nucleotide data (i.e., whereby synonymous substitutions are eliminated by ambiguity recoding) do not support either liverworts or mosses as the earliest-branching land plants (Supplementary Fig. S17). Partitioning deg-nt data in ML analyses has no obvious effect on tree topologies (data not shown). The model gene-1-2-3 is favored by both AIC and BIC (Supplementary Table S3 and Fig. S17). The hypothesis that liverworts compose the sister-group to the remaining land plant lineages received only 46–60% bootstrap supports in six different partitioning schemes. However, the common ancestry between hornworts and tracheophytes is always strongly supported in $ML_{deg\text{-}nt}$

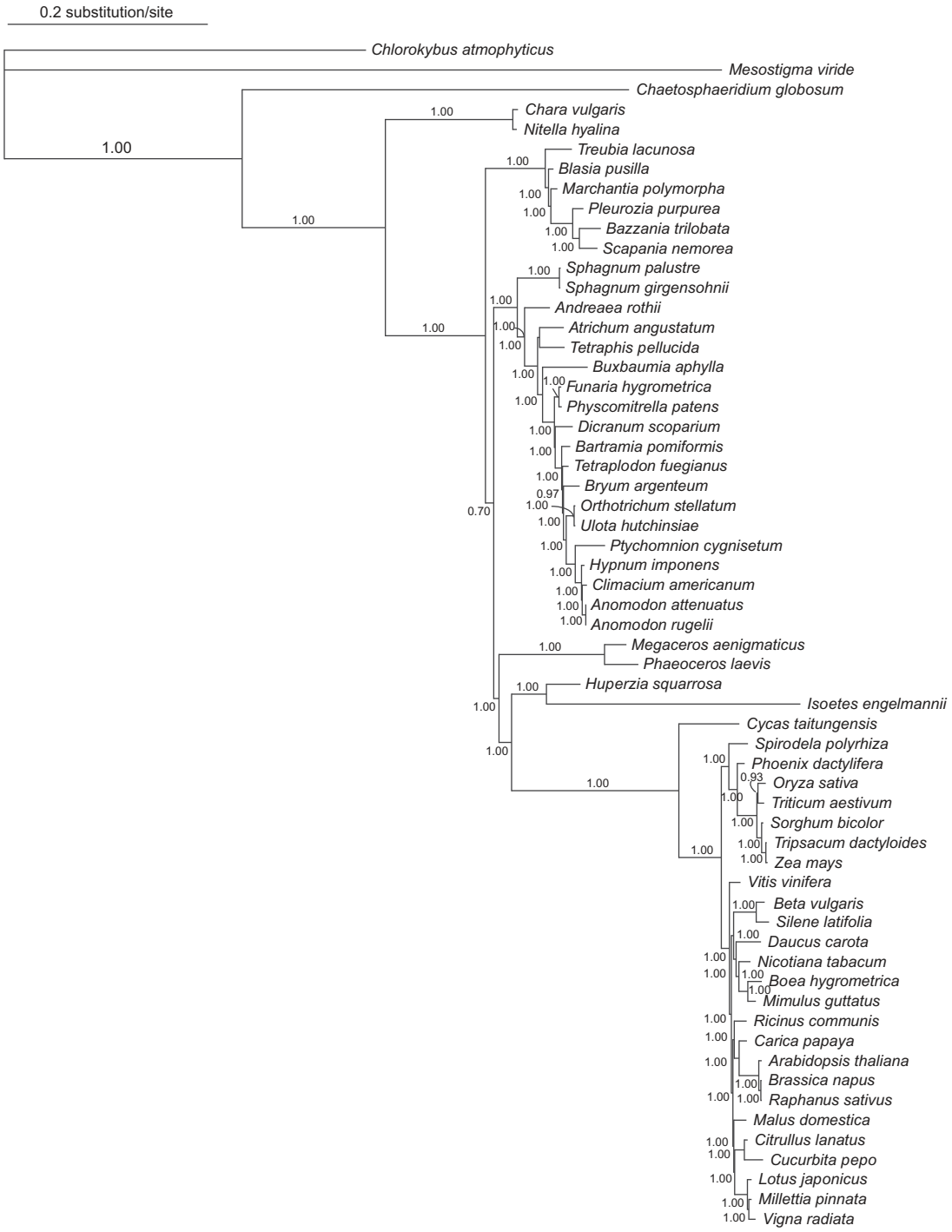

FIGURE 3. Majority-rule consensus tree of a composition-heterogeneous Bayesian analysis of the nt data. Nonstationary compositional MCMC were performed in P4 using duplicate runs each consisting of 2 million generations using a GTR + I + Γ model with two composition vectors (CV2) and including a polytomy prior (PP). Details of the chain diagnostics can be found in Supplementary Figure S11.

analyses (BS = 93–100%). Similarly, liverworts rooting the deepest in the land plant tree did not receive significant support under Bayesian reconstructions on codon-degenerated data ($BI_{deg\text{-}nt}$-PP = 0.85–0.88; Supplementary Figs. S18 and S19); however, the hornwort–tracheophyte sister-group relationship is maximally supported in the same trees.

Alternative Hypotheses Testing

Constrained nonparametric bootstrapping analyses of both nucleotide and amino acid data sets consistently rejected three alternative hypotheses: liverworts and mosses forming a clade, bryophytes monophyletic, and hornworts being the earliest land plant lineage. These tests did, however, not reject the hypotheses of liverworts

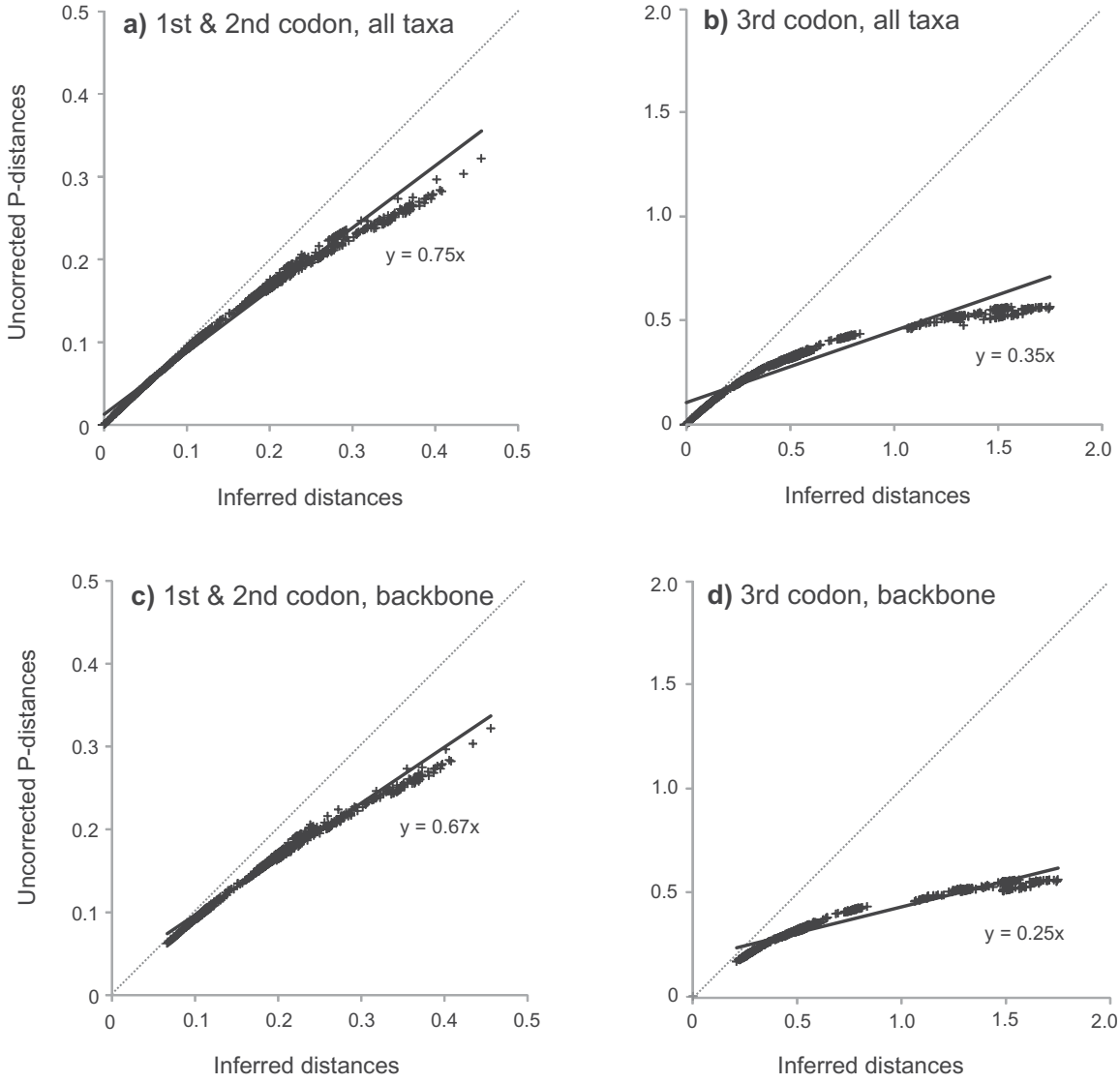

FIGURE 4. Saturation level in the first + second codon positions, and the third codon positions. The uncorrected *P*-distances were plotted against the inferred distances. The *x*-axis corresponds to the distance inferred by the maximum-likelihood method using the Tamura–Nei model (Tamura and Nei, 1993), and *y*-axis corresponds to the uncorrected distance observed for the same taxa pair. a) and b) include pairwise comparisons between all exemplars across the whole tree. c) and d) include only comparisons between two taxa drawn from two of the five major lineages (A, L, M, H, or V) of the backbone phylogeny. The level of saturation was estimated by computing the slope of the regression line, the lesser the slope, the greater the level of saturation.

or mosses sister to the remaining land plants when inferences are drawn from nucleotide or amino acid data, respectively (Table 1). By contrast, all alternative phylogenetic hypotheses can be rejected ($P \leq 0.05$) using the more powerful parametric bootstrapping method (the SOWH test).

DISCUSSION

Suitability of mt Data for Resolving Early Land Plant Phylogeny

Within the currently accepted phylogeny of land plants (Qiu 2008), bryophytes compose a grade marking the successful transition to land and leading to the origin of vascular plants. This phylogenetic hypothesis is, however, incongruent with previous inferences and not unambiguously supported by concatenated data from three genomic compartments (Qiu et al. 2006b). Conflicting topologies usually lack strong support for the deep nodes likely due to insufficient data (e.g., Hori et al. 1985; Malek et al. 1996; Duff and Nickrent 1999). Highly supported phylogenies were inferred from genomic data but may not be reliable because of limited taxon, and in particular bryophyte, sampling (Nishiyama et al. 2004; Chaw et al. 2008). The mt genomic data assembled here support the paraphyly of bryophytes, and the uniquely shared ancestry between

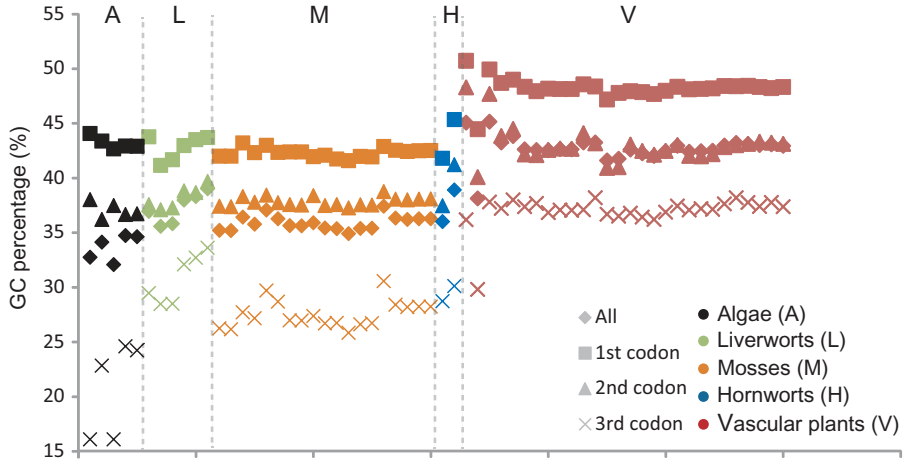

FIGURE 5. Nucleotide GC contents percentage of each species for the entire set of genes, and first, second, or third codon positions only. Taxa are sorted on the *x*-axis in the same order as on the phylogeny (Fig. 3).

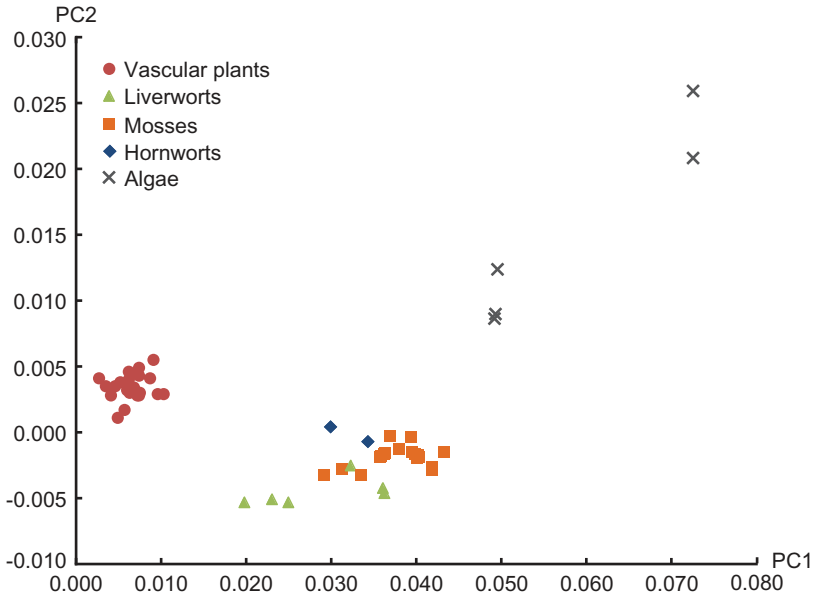

FIGURE 6. RSCU correspondence analyses of each land plant lineage.

hornworts and vascular plants, which is more strongly supported (ML_{aa}-BS = 98%, $BI_{aa\&nt}$-PP = 1.0; Fig. 2a,b) than when inferred from limited character or taxon samples (Qiu et al. 2006b, 2007; Gao et al. 2010; Chang and Graham 2011). A closer relationship of hornworts to vascular plants than to other bryophytes is consistently supported by the four-cluster (i.e., four taxon) likelihood mapping analyses, and also by inferences from the nt or aa data from only the 20 genes present in hornworts (results not shown), indicating that this relationships is not biased by taxon sampling or missing data. Furthermore, the data reject the hypotheses of a monophyletic bryophyte lineage (Hori et al. 1985; Nishiyama et al. 2004; Cox et al. 2014) or sister-group relationship between mosses and liverworts (Nishiyama and Kato 1999; Nickrent et al. 2000; Renzaglia et al. 2000) (Table 1). The phylogenetic signal of the mt exons for such deep nodes is strong, but the relationships of mosses and liverworts remain ambiguous due to conflicting signals in nt data and their aa translation.

Plant mt genes may transfer horizontally (HGT) between unrelated organisms (Bergthorsson et al. 2003; Won and Renner 2003) and then undergo pseudogenization (Bergthorsson et al. 2004; Rice et al. 2013), and consequently mislead phylogenetic reconstructions. We can reject the hypothesis of HGT among lineages of land plants affecting our inferences because we drew phylogenetic information only from clearly functional mt genes and because inferences from each individual locus recovered the main lineages (i.e., liverworts, mosses, hornworts, and tracheophytes) as monophyletic (Fig. 1).

TABLE 1. Comparison between the optimal ML tree and the alternative hypotheses

Data	Optimal topology	Alternative hypothesis	Nonparametric bootstrapping tests					Parametric bootstrapping SOWH test	
			Δ log *L*	S.E.	SH	KH	AU	Δ log *L* simulated ($P \leq 0.05$)	Δ log *L* observed
nt	M,(L,(H,V))	L,(M,H,V)	8.48	14.25	0.46	0.28	0.28	0.01	6.93+
		(L,M)	27.25	12.29	0.02+	0.01+	0.01+	0.01	15.78+
		(L,M,H)	47.20	16.30	0.00+	0.00+	0.01+	0.14	22.08+
		H,(L,M,V)	47.36	16.17	0.00+	0.00+	0.00+	2.11	24.48+
Aa	L,(M,(H,V))	M,(L,H,V)	11.65	10.60	0.32	0.14	0.15	0.00	8.64+
		(L,M)	16.17	9.35	0.05+	0.05+	0.05+	0.00	10.58+
		(L,M,H)	52.93	15.57	0.00+	0.00+	0.00+	3.57	26.08+
		H,(L,M,V)	48.79	17.42	0.00+	0.00+	0.00+	0.00	22.12+

Notes: In nonparametric bootstrapping analyses, *P*-values were estimated by Consel. The sign "+"indicates the alternative topology differing significantly (*P*-value at 0.05) from the ML tree, and should be rejected. In parametric bootstrapping SOWH test, the Δ log *L* obtained from unconstraint (optimal) and constraint (alternative hypothesis) analyses of original data set was compared with the Δ log *L* (*P*-value at 0.05) from simulated data set, if the former is lager, the alternative hypothesis is rejected. Abbreviations: L, liverworts; M, mosses; H, hornworts; and V, vascular plants.

Incongruence between nt- and aa-Based Trees and the Possible Causes

The standard homogeneous analyses based on nucleotide and amino acid data led to optimal trees that differ in the placement of mosses or liverworts, respectively, as the most deeply rooted embryophyte lineage, in each case with maximum PP (BI-PP = 1.0; Fig. 2a,b). Support for the conflicting relationships is, by contrast, only moderate under the ML criterion (ML_{nt}-BS = 70%, ML_{aa}-BS = 82%; Fig. 2a,b), but the parametric bootstrapping SOWH test under the homogeneous model rejects all alternative hypotheses for each data set. This suggests that the phylogenetic signal in both the nucleotide and amino acid data set is strong, and hence that the conflict is significant. Compared with the SOWH test, the nonparametric bootstrapping (SH, KH, and AU) tests failed to reject the competing hypotheses in both nucleotide and amino acid data (Table 1). The SOWH test is considered stricter than nonparametric tests in rejecting a null hypothesis, and hence is less prone to Type I error (Buckley 2002). The nonparametric bootstrapping method relies on simpler models, and this may account for the difference in assessing the strength of the phylogenetic signal. The nt data supported mosses as the earliest diverging lineage among extant land plants: a tree that is in direct conflict with the currently accepted hypothesis inferred from concatenated loci (Qiu et al. 2006b, 2007) or chloroplast genomic data (Wolf et al. 2005; Gao et al. 2010; Chang and Graham 2011), which resolve liverworts as the sister-group to the remaining embryophytes. The hypothesis that mosses diverged first as inferred from the nt data is novel, and is thus in conflict with the current widely adopted view of liverwort marking the transition to land. To resolve this incongruence, we assessed the nature of the support for either hypothesis, by contrasting the optimal inferences inferred under distinct models or subsets of the data.

Substitutional saturation decreases the accuracy of phylogenetic signal, and affects nucleotide data more rapidly than amino acid data, due to the redundancy of the genetic code. Saturation is a function of time and mutation rate and hence affects primarily the resolution of deep nodes as shown in the reconstruction of deep arthropod (Xia et al. 2003) or animal phylogeny (Philippe et al. 2011). When sequences have experienced full saturation, the similarity between the sequences may depend only on the nucleotide frequencies, so that phylogenetic inferences based on such data will reflect artifactual relationships (Xia 2000) based on similarities in composition among sites and among lineages. In phylogenetic analyses, substitution models can account for saturation only to a certain extent by modeling site rates (typically a mixture model of gamma-distributed among-site rates) and modeling compositional biases (i.e., CAT and NDCH models). The common practice to avoid problems associated with saturation in protein-coding data sets is to remove from protein-coding nt data sets the third codon positions, which are mostly redundant in defining the codon and hence tend to have the fastest substitution rate. Alternatively, since many substitutions in protein-coding genes yield synonymous codons, translating the nt sequence into an aa sequence can reduce the negative phylogenetic effects of saturation at the nucleotide level, too. Within our data set, the third codon positions are more heavily saturated than first and second codon positions, especially when deeply rooted exemplars are compared (Fig. 4d), and such marked differences may underlay the conflict in phylogenetic signal of the two subsets of data.

On average, third codon positions may be more saturated than sites in positions first or second, but not all third positions have undergone multiple substitutions, whereas some first and second codon positions may have become saturated. Therefore, assuming that all third codon positions alone are the source of homoplasy is an oversimplification, resulting in nonsaturated third positions being excluded and saturated first and second position being retained. We tested the effect of such potentially imperfect site partitioning by identifying fast sites regardless of their codon position, and removing them incrementally from the nt alignment using the

observed variability (OV)-sorting method (Goremykin et al. 2010). ML analyses (results not shown) showed that removing 3% of the fastest sites still yielded mosses sister to the remaining land plants, but removing 7% or more of the fastest sites resulted in liverworts as the earliest diverging lineage. This result is consistent with our degenerate-recoding analyses, which also yielded an early-branching liverworts topology because the majority of the fastest OV sites are likely those experiencing the highest number of synonymous substitutions—7% site removal included 49% third, 22% second, and 28% first positions. Moreover, the result is also consistent with our NDCH analyses where modeling among-lineage composition heterogeneity also resulted in an early-branching liverworts topology suggesting that fastest OV-scored sites are responsible for a compositional bias among lineages in analyses using composition stationary models that result in the mosses diverging first. Removing 23% of the fast-evolving sites led to a collapsed tree, and removing 13%, 17%, and 20% of the fastest sites yielded hornworts sister to lycophytes only, a relationship that is highly unlikely to be correct. These observations suggest that by removing all third codon positions, the main source of the conflict was deleted from the data set.

Another source of incongruence in phylogenetic signal of nucleotide and amino acid data sets may be a lineage-specific base compositional heterogeneity, which is known to cause artifacts when not accommodated in the models (Lockhart et al. 1992, 1994; Mooers and Holmes 2000; Foster 2004). Nonstationary heterogeneous composition models, which allow base composition to vary among lineages, may prevent such topological artifacts (Yang and Roberts 1995; Foster 2004). In the current mt data set, analyses of the GC percentage plot (Fig. 5) and the *T*-tests (Supplementary Table S4) indicated that major lineages of land plants significantly differ in their GC composition. Inference from nt data using nonstationary composition model that allow GC composition to vary among lineages recovers liverworts as the earliest land plant lineage (Fig. 3) suggesting that compositional biases may also underlay the incongruence between inferences from stationary composition models and aa data.

Among-site composition heterogeneity in the alignment can also bias phylogenetic inference (Lartillot and Philippe 2004). However, Bayesian MCMC analyses invoking the composition site-heterogeneous CAT–GTR model also resolved mosses sister to the remaining land plants and so did composition site-homogeneous models when inferences are drawn from nt data. In contrast, analyses of aa data using the same CAT–GTR model did not support the liverworts as sister to the rest of land plants, as did site-composition homogeneous models, although the hornworts were strongly supported as the sister-group to tracheophytes. This observation implies that the strong support for the early branching of liverworts when using homogeneous models may be due in part to variation in composition among sites in the amino acid data.

Degeneracy in the universal genetic code, where multiple codons specifying a single type of amino acid, enables synonymous codons to be used differentially within a genome and across species, potentially resulting in highly distinct patterns of codon-usage among lineages (Liu et al. 2004; Zhou and Li 2009; Wang et al. 2010; Plotkin and Kudla 2011). Among lineages codon-usage bias can result in homoplasious character distributions and thus artifactual phylogenetic inferences (Foster 2004; Inagaki et al. 2004; Inagaki and Roger 2006; Regier et al. 2010; Rota-Stabelli et al. 2013; Cox et al. 2014). As most substitutions in third codon positions are synonymous, exclusion of these positions can partially resolve the problem of compositional convergence, but not sufficiently to remove all the influences due to codon degeneracy at the first or second codon position. To assess and lessen the effect of convergent compositional heterogeneity facilitated by codon redundancy, nucleotide sequences can be recoded to a sequence of codons integrating the degeneracy of the genetic code through the use of IUPAC codes reflecting alternative "synonymous" bases at the three codon positions (Criscuolo and Gribaldo 2010; Regier et al. 2010). The reconstruction of the early cladogenic events from sequences degenerated to ambiguous codons reflecting codon synonymy led to a hypothesis congruent with that inferred from aa data, namely liverworts composing the sister-group to the remaining land plants (albeit without statistical support), suggesting that among-lineage codon usage and/or compositional biases may also contribute to the nt–aa conflict. A similar result was observed in arthropods (Regier et al. 2010; Rota-Stabelli et al. 2013), dinoflagellates (Inagaki et al. 2004), and red algae (Inagaki and Roger 2006). Although mosses and liverworts exhibit similar codon usages, the usage of mosses are slightly more similar to those expressed in algae (Fig. 6), a trend that may introduce a phylogenetic bias in nt data (see also Cox et al. 2014) whereby mosses, rather than liverworts compose the least-derived macroevolutionary lineage of land plants. Recoding the codons in nt data to reflect the degeneracy of the genetic code is expected to decrease the total amount of phylogenetic signal in the data by replacing specific nucleotides with ambiguity codes accommodating synonymous codons. Bootstrap support values from deg-nt analyses are, indeed, generally lower (Supplementary Fig. S18) compared with the original nt data (Supplementary Fig. S17). However, degenerating nt data by codons to eliminate "noisy" phylogenetic signals from the codon/composition biases may indeed allow an underlying, and accurate, phylogenetic signal to dominate. Codon-usage bias is correlated with GC compositional bias (Stenøien 2005). Even in the absence of selective constraints, variation in patterns of neutral processes such as directional mutation bias can result in convergence of base composition, and thereby of dependent codon-usage biases, because higher GC composition leads to selection of GC rich codons (Sharp et al. 2010). But it is

not clear whether shifts in GC composition determine synonymous codon usage or whether biases for specific codons shape the GC composition (Knight et al. 2001; Sharp et al. 2010; Behura and Severson 2013).

Recent phylogenetic analyses using genomic-scale mtDNA or cpDNA data, focusing on land plants (Chaw et al. 2008; Chang and Graham 2011), ferns (Rai and Graham 2010), angiosperms (Jansen et al., 2007), or broad sampling of one group of angiosperms (Moore et al. 2010; Barrett et al. 2012) failed to test the potential bias introduced by the above parameters shaping the evolution of individual loci. We suggest that inferences from genomic data should always assess the effect of saturation, and composition and codon-usage heterogeneity—processes whose effects are not independent but potentially correlated—particularly if the divergences of interest are ancient (e.g., origin of land plants) or the genomes evolving fast (e.g., nuclear exomes).

Conclusions

Although phylogenomic analyses can produce robust trees, the well-supported results may still be erroneous. Reconstruction of land plant backbone phylogeny using entire mt exomes yields incongruent hypotheses when inferences are drawn from nt data versus from their aa translations. Correcting the nt data for systematic errors, such as saturation and among-lineage codon usage and compositional heterogeneity biases recovered a topology congruent with inferences from the aa data, whereby bryophytes are paraphyletic, with liverworts sister to the other land plants (but not significantly supported) and hornworts sister to vascular plants. In addition, among-site composition analyses of the amino acid data also question the early divergence of the liverworts; hence, we are unable to draw any firm conclusion concerning the earliest-diverging land plant lineage from the current data. In contrast, all current analyses suggest hornworts as the sister-group to the tracheophytes indicating the paraphyly of bryophytes with mt data and seemingly in direct conflict with recent analyses of plastid data that show bryophyte monophyly (Cox et al. 2014). We note that both the algal Zygnematales—most likely the sister-group to land plants (Wodniok et al. 2011; Civáň et al. 2014)—and the ferns are missing from the current analyses and that their inclusion in future data sets may have a significant impact on the resolution of the relationships among land plant lineages.

With the development of high-throughput sequencing techniques, plant mt genomes can be effectively and efficiently sequenced, composing an important resource for plant phylogenetic reconstructions. Land plants likely went through a rapid radiation at an early stage (Kenrick and Crane 1997). During the rapid diversification, the speciation events are closely spaced in time, and thus the amount of phylogenetic signal is often small, as exemplified perhaps by the short branches marking deep cladogenic events following the transition to land. Furthermore, because the events are ancient, the relatively scarce phylogenetic signals can be out-weighted by homoplasy driven by saturation or heterogeneity in nucleotide compositions among species, especially when relationships are reconstructed from primary nucleotide data. We therefore suggest that phylogenetic hypotheses inferred from genomic-scale nucleotide data be interpreted with caution, even when highly supported, until the effects of parameters linked to redundancy of the genetic code are fully assessed.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://dx.doi.org/10.5061/dryad.7b470.

Funding

This work was supported by the grants from the National Science Foundation [EF 0531557, DEB 0919284, and DEB-1146295 to B.G.].

Acknowledgements

The authors thank Vincent Savolainen, Harald Schneider, Frank Anderson, and an anonymous reviewer for their constructive comments and suggestions. They would like to thank Paul O. Lewis (UConn) for discussion on the article, and help with the SOWH test. They thank Laura L. Forrest and Lily R. Lewis (UConn) for sharing the genomic data of *Blasia* and *Tetraplodon*, and Virge Kask (UConn) for making illustrations in Figure 2. They also thank University of Connecticut Health Center Translational Genomics Core Facility (Farmington, CT, USA) for use of the Illumina instrument. The UConn Bioinformatics Facility and the Plant Systematics and Bioinformatics Research Group (PSB) at CCMAR provided computing resources for the Bayesian- and maximum-likelihood phylogenetic analyses performed for this study. Blaise Li (CCMAR) provided a Python software implementation of the OV scoring procedure.

References

Abascal F., Zardoya R., Telford M.J. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38:W7–W13.

Adachi J., Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42:459–468.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19:716–723.

Barrett C.F., Davis J.I., Leebens-Mack J., Conran J.G., Stevenson D.W. 2012. Plastid genomes and deep relationships among the commelinid monocot angiosperms. Cladistics 29:65–87.

Bateman R.M., Crane P.R., DiMichele W.A., Kenrick P.R., Rowe N.P., Speck T., Stein W.E. 1998. Early evolution of land plants: phylogeny, physiology, and ecology of the primary terrestrial radiation. Annu. Rev. Ecol. Syst. 29:263–292.

Behura S.K., Severson D.W. 2013. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol. Rev. 88:49–61.
Bergthorsson U., Adams K.L., Thomason B., Palmer J.D. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424:197–201.
Bergthorsson U., Richardson A.O., Young G.J., Goertzen L.R., Palmer J.D. 2004. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm Amborella. Proc. Natl Acad. Sci. U. S. A. 101:17747–17752.
Blanquart S., Lartillot N. 2008. A site-and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25:842–858.
Bremer K. 1985. Summary of green plant phylogeny and classification. Cladistics 1:369–385.
Brown J.M., Lemmon A.R. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.
Buckley T.R. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst. Biol. 51:509–523.
Chang Y., Graham S.W. 2011. Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. Am. J. Bot. 98:839–849.
Chaw S.M., Shih A.C., Wang D., Wu Y.W., Liu S.M., Chou T.Y. 2008. The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol. Biol. Evol. 25:603–615.
Civáň P., Foster P.G., Embley T.M., Séneca A., Cox C.J. 2014. Analyses of charophyte chloroplast genomes help characterize the ancestral chloroplast genome of land plants. Genome Biol. Evol. 6: 897–911.
Cox C.J., Foster P.G., Hirt R.P., Harris S.R., Embley T.M. 2008. The archaebacterial origin of eukaryotes. Proc. Natl Acad. Sci. U. S. A. 105:20356–20361.
Cox C.J., Li B., Foster P.G., Embley T.M., Civáň P. 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. Syst. Biol. 63:272–279.
Crandall-Stotler B., Stotler R.E., Long D.G. 2009. Morphology and classification of the Marchantiophyta. In: Goffinet B., Shaw A.J., editors. Bryophyte biology. 2nd ed. Cambridge (UK): Cambridge University Press. p. 1–54.
Criscuolo A., Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10:210.
Dang C.C., Lefort V., Le V.S., Le Q.S., Gascuel O. 2011. ReplacementMatrix: a web server for maximum-likelihood estimation of amino acid replacement rate matrices. Bioinformatics 27:2758–2760.
Darling A.C.E., Mau B., Blattner F.R., Perna N.T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394–1403.
Delsuc F., Brinkmann H., Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.
Drouin G., Daoud H., Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol. Phylogenet. Evol. 49:827–831.
Duff R.J., Nickrent D.L. 1999. Phylogenetic relationships of land plants using mitochondrial small-subunit rDNA sequences. Am. J. Bot. 86:372–386.
Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.
Edwards D., Morris J.L., Richardson J.B., Kenrick P. 2014. Cryptospores and cryptophytes reveal hidden diversity in early land floras. New Phytol. 202:50–78.
Finet C., Timme R.E., Delwiche C.F., Marlétaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr. Biol. 20:2217–2222.
Finster S., Legen J., Qu Y., Schmitz-Linneweber C. 2012. Land plant RNA editing or: don't be fooled by plant organellar DNA sequences. In: Bock R., Knoop V., editors. Genomics of chloroplasts and mitochondria. Dordrecht: Springer. p. 293–321.
Fletcher W., Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. Mol. Biol. Evol. 26:1879–1888.
Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.
Gao L., Su Y.J., Wang T. 2010. Plastid genome sequencing, comparative genomics, and phylogenomics: current status and prospects. J. Syst. Evol. 48:77–93.
Garbary D.J., Renzaglia K. 1998. Bryophyte phylogeny and the evolution of land plants: evidence from development and ultrastructure. In: Bates J.W., Ashton N.W., Duckett J.G., editors. Bryology for the twenty-first century. Leeds (UK): Maney and the British Bryological Society. p. 45–63.
Garbary D.J., Renzaglia K.S., Duckett J.G. 1993. The phylogeny of land plants: a cladistic analysis based on male gametogenesis. Plant Syst. Evol. 188:237–269.
Goffinet B. 2000. Origin and phylogenetic relationships of bryophytes. In: Shaw A.J., Goffinet B., editors. The biology of bryophytes. Cambridge (UK): Cambridge University Press. p. 124–149.
Goffinet B., Buck W.R. 2013. The evolution of body form in bryophytes. Ann. Plant Rev. 45:51–90.
Goffinet B., Buck W.R., Shaw A.J. 2009. Morphology, anatomy, and classification of the Bryophyta. In: Goffinet B., Shaw A.J., editors. Bryophyte biology. 2nd ed. Cambridge (UK): Cambridge University Press. p. 55–138.
Goldman N., Anderson J.P., Rodrigo A.G. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.
Goremykin V.V., Nikiforova S.V., Bininda-Emonds O.R. 2010. Automated removal of noisy data in phylogenomic analyses. J. Mol. Evol. 71:319–331.
Gouy M., Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10:7055–7074.
Graur D., Li W.-H. 2000. Fundamentals of molecular evolution. 2nd ed. Sunderland (MA): Sinauer Associates.
Hedderson T.A., Chapman R.L., Rootes W. 1996. Phylogenetic relationships of bryophytes inferred from nuclear-encoded rRNA gene sequences. Plant Syst. Evol. 200:213–224.
Hori H., Lim B.-L., Osawa S. 1985. Evolution of green plants as deduced from 5S rRNA sequences. Proc. Natl Acad. Sci. U. S. A. 82:820–823.
Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.
Huson D.H., Dezulian T., Klopper T., Steel M.A. 2004. Phylogenetic super-networks from partial trees. IEEE Trans. Comput. Biol. Bioinform. 1:151–158.
Inagaki Y., Roger A.J. 2006. Phylogenetic estimation under codon models can be biased by codon usage heterogeneity. Mol. Phylogenet. Evol. 40:428–434.
Inagaki Y., Simpson A.G., Dacks J.B., Roger A.J. 2004. Phylogenetic artifacts can be caused by leucine, serine, and arginine codon usage heterogeneity: dinoflagellate plastid origins as a case study. Syst. Biol. 53:582–593.
Jansen R.K., Cai Z., Raubeson K., Daniell H., Depamphilis C.W., Leebens-Mack J., Müller K.F., Guisinger-Bellian M., Haberle R.C., Hansen A.K. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl Acad. Sci. U. S. A. 104:19369–19374.
Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.
Jian S., Soltis P.S., Gitzendanner M.A., Moore M.J., Li R.Q., Hendry T.A., Qiu Y.L., Dhingra A., Bell C.D., Soltis D.E. 2008. Resolving an ancient, rapid radiation in Saxifragales. Syst. Biol. 57:38–57.
Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.
Karol K.G., Arumuganathan K., Boore J.L., Duffy A.M., Everett K.D.E., Hall J.D., Hansen S.K., Kuehl J.V., Mandoli D.F., Mishler B.D. 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.
Kass R.E., Raftery A.E. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795.
Katoh K., Kuma K.-I., Toh H., Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Keane T.M., Creevey C.J., Pentony M.M., Naughton T.J., Mclnerney J.O. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol. Biol. 6:29.

Kelch D.G., Driskell A., Mishler B.D. 2004. Inferring phylogeny using genomic characters: a case study using land plant plastomes. In: Goffinet B, Hollowell V, Magill R, editors. Molecular systematics of bryophytes. St. Louis: Missouri Botanical Garden Press. p. 3–12.

Kenrick P., Crane P.R. 1997. The origin and early diversification of land plants. A cladistic study. Washington (DC): Smithsonian Institute Press.

Kenrick P., Wellman C.H., Schneider H., Edgecombe G.D. 2012. A timeline for terrestrialization: consequences for the carbon cycle in the Palaeozoic. Phil. Trans. R. Soc. Lond. B Biol. Sci. 367:519–536.

Kishino H., Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.

Knight R.D., Freeland S.J., Landweber L.F. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2:research0010.

Knoop V. 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr. Genet. 46:123–139.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Lartillot N., Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.

Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lewis L.A., Mishler B.D., Vilgalys R. 1997. Phylogenetic relationships of the liverworts (Hepaticae), a basal embryophyte lineage, inferred from nucleotide sequence data of the chloroplast gene *rbc*L. Mol. Phylogenet. Evol. 7:377–393.

Lewis P.O., Holder M.T., Holsinger K.E. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253.

Li C., Lu G., Orti G. 2008. Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Syst. Biol. 57:519–539.

Liu Q., Feng Y., Xue Q. 2004. Analysis of factors shaping codon usage in the mitochondrion genome of *Oryza sativa*. Mitochondrion 4:313–320.

Liu Y., Xue J.Y., Wang B., Li L., Qiu Y.L. 2011. The mitochondrial genomes of the early land plants *Treubia lacunosa* and *Anomodon rugelii*: dynamic and conservative evolution. PLoS One 6:e25836.

Liu Y., Forrest L.L., Bainard J.D., Budke J.M., Goffinet B. 2013. Organellar genome, nuclear ribosomal DNA repeat unit, and microsatellites isolated from a small-scale of 454 GS FLX sequencing on two mosses. Mol. Phylogenet. Evol. 66:1089–1094.

Lockhart P.J., Howe C.J., Bryant D.A., Beanland T.J., Larkum A.W.D. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. 34:153–162.

Lockhart P.J., Steel M.A., Hendy M.D., Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605–612.

Maddison W.P., Maddison D. 2001. Mesquite: a modular system for evolutionary analysis. Version 2.74. Available from: URL http://mesquiteproject.org.

Malek O., Lättig K., Hiesel R., Brennicke A., Knoop V. 1996. RNA editing in bryophytes and a molecular phylogeny of land plants. EMBO J. 15:1403–1411.

McGuire J.A., Witt C.C., Altshuler D.L., Remsen J. 2007. Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Syst. Biol. 56:837–856.

McInerney J.O. 1998. GCUA: general codon usage analysis. Bioinformatics 14:372–373.

Mishler B.D., Churchill S.P. 1984. A cladistic approach to the phylogeny of the "bryophytes". Brittonia 36:406–424.

Mooers A.Ø., Holmes E.C. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15:365–369.

Moore M.J., Soltis P.S., Bell C.D., Burleigh J.G., Soltis D.E. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc. Natl Acad. Sci. U. S. A. 107:4623–4628.

Newton M.A., Raftery A.E. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Statist. Soc. Ser. B 56:3–48.

Nickrent D.L., Parkinson C.L., Palmer J.D., Duff R.J. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Mol. Biol. Evol. 17:1885–1895.

Nishiyama T., Kato M. 1999. Molecular phylogenetic analysis among bryophytes and tracheophytes based on combined data of plastid coded genes and the 18S rRNA gene. Mol. Biol. Evol. 16:1027–1036.

Nishiyama T., Wolf P.G., Kugita M., Sinclair R.B., Sugita M., Sugiura C., Wakasugi T., Yamada K., Yoshinaga K., Yamaguchi K. 2004. Chloroplast phylogeny indicates that bryophytes are monophyletic. Mol. Biol. Evol. 21:1813–1819.

Oda K., Yamato K., Ohta E., Nakamura Y., Takemura M., Nozato N., Akashi K., Kanegae T., Ogura Y., Kohchi T., Ohyama, K. 1992. Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA—a: A primitive form of plant mitochondrial genome. J. Mol. Biol. 223:1–7.

Palmer J.D., Herbon L.A. 1988. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. J. Mol. Evol. 28:87–97.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Philippe H., Forterre P. 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. 49:509–523.

Plotkin J.B., Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12:32–42.

Qiu Y.L. 2008. Phylogeny and evolution of charophytic algae and land plants. J. Syst. Evol. 46:287–306.

Qiu Y.L., Cho Y., Cox J.C., Palmer J.D. 1998. The gain of three mitochondrial introns identifies liverworts as the earliest land plants. Nature 394:671–674.

Qiu Y.L., Li L., Hendry T.A., Li R.Q., Taylor D.W., Issa M.J., Ronen A.J., Vekaria M.L., White A.M. 2006a. Reconstructing the basal angiosperm phylogeny: evaluating information content of mitochondrial genes. Taxon 55:837–856.

Qiu Y.L., Li L., Wang B., Chen Z.D., Knoop V., Groth-Malonek M., Dombrovska O., Lee J., Kent L., Rest J. 2006b. The deepest divergences in land plants inferred from phylogenomic evidence. Proc. Natl Acad. Sci. U. S. A. 103:15511–15516.

Qiu Y.L., Li L., Wang B., Chen Z.D., Dombrovska O., Lee J., Kent L., Li R.Q., Jobson R.W., Hendry T.A. 2007. A nonflowering land plant phylogeny inferred from nucleotide sequences of seven chloroplast, mitochondrial, and nuclear genes. Int. J. Plant Sci. 168:691–708.

Qiu Y.L., Li L., Wang B., Xue J.Y., Hendry T.A., Li R.Q., Brown J.W., Liu Y., Hudson G.T., Chen Z.D. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. J. Syst. Evol. 48:391–425.

Rai H.S., Graham S.W. 2010. Utility of a large, multigene plastid data set in inferring higher-order relationships in ferns and relatives (monilophytes). Am. J. Bot. 97:1444–1456.

Rambaut A., Grass N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Regier J.C., Shultz J.W., Zwick A., Hussey A., Ball B., Wetzer R., Martin J.W., Cunningham C.W. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463:1079–1083.

Renzaglia K.S., Duff R.J., Nickrent D.L., Garbary D.J. 2000. Vegetative and reproductive innovations of early land plants: implications for a unified phylogeny. Phil. Trans. R. Soc. Lond. B Biol. Sci. 355:769–793.

Renzaglia K.S., Villarreal J.C., Duff R.J. 2009. New insights into morphology, anatomy, and systematics of hornworts. In: Goffinet B., Shaw A.J., editors. Bryophyte biology. 2nd ed. Cambridge (UK): Cambridge University Press. p. 139–171.

Rice D.W., Alverson A.J., Richardson A.O., Young G.J., Sanchez-Puerta M.V., Munzinger J., Barry K., Boore J.L., Zhang Y., Knox E.B. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. Science 342:1468–1473.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Rota-Stabelli O., Lartillot N., Philippe H., Pisani D. 2013. Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. Syst. Biol. 62:121–133.

Ruhfel B.R., Gitzendanner M.A., Soltis P.S., Soltis D.E., Burleigh J.G. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol. Biol. 14:23.

Schmidt H.A., Strimmer K., Vingron M., von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.

Schwarz G. 1978. Estimating the dimension of a model. Ann. Statist. 6:461–464.

Sharp P.M., Emery L.R., Zeng K. 2010. Forces that influence the evolution of codon bias. Phil. Trans. R. Soc. Lond. B Biol. Sci. 365:1203–1212.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

Shimodaira H., Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stenøien H.K. 2005. Adaptive basis of codon usage in the haploid moss *Physcomitrella patens*. Heredity 94:87–93.

Strimmer K., von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl Acad. Sci. U. S. A. 94:6815–6819.

Swofford D.L. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4.0 b10. Sunderland (MA): Sinauer Associates.

Talavera G., Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564–577.

Tamura K., Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.

Taylor T.N., Kerp H., Hass H. 2005. Life history biology of early land plants: deciphering the gametophyte phase. Proc. Natl Acad. Sci. U. S. A. 102:5892–5897.

Terasawa K., Odahara M., Kabeya Y., Kikugawa T., Sekine Y., Fujiwara M., Sato N. 2007. The mitochondrial genome of the moss *Physcomitrella patens* sheds new light on mitochondrial evolution in land plants. Mol. Biol. Evol. 24:699–709.

Turmel M., Otis C., Lemieux C. 2013. Tracing the evolution of streptophyte algae and their mitochondrial genome. Genome Biol. Evol. 5:1817–1835.

Wang B., Xue J., Li L., Liu Y., Qiu Y.L. 2009. The complete mitochondrial genome sequence of the liverwort Pleurozia purpurea reveals extremely conservative mitochondrial genome evolution in liverworts. Curr. Genet. 55:601–609.

Wang B., Liu J., Jin L., Feng X.Y., Chen J.Q. 2010. Complex mutation and weak selection together determined the codon usage bias in bryophyte mitochondrial genomes. J. Integr. Plant Biol. 52:1100–1108.

Whitfield J.B., Lockhart P.J. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.

Wodniok S., Brinkmann H., Glöckner G., Heidel A.J., Philippe H., Melkonian M., Becker B. 2011. Origin of land plants: do conjugating green algae hold the key? BMC Evol. Biol. 11:104.

Wolf P.G., Karol K.G., Mandoli D.F., Kuehl J., Arumuganathan K., Ellis M.W., Mishler B.D., Kelch D.G., Olmstead R.G., Boore J.L. 2005. The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). Gene 350:117–128.

Wolfe K.H., Li W.H., Sharp P.M. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl Acad. Sci. U. S. A. 84:9054–9058.

Won H., Renner S.S. 2003. Horizontal gene transfer from flowering plants to Gnetum. Proc. Natl Acad. Sci. U. S. A. 100:10824–10829.

Xia X. 2000. Data analysis in molecular biology and evolution. Boston (MA): Kluwer Academic Publishers.

Xia X., Xie Z., Salemi M., Chen L., Wang Y. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26:1–7.

Yang Z., Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12:451–458.

Zhang T., Zhang X., Hu S., Yu J. 2011. An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. Plant Methods 7:38.

Zhong B., Liu L., Yan Z., Penny D. 2013. Origin of land plants using the multispecies coalescent model. Trends Plant Sci. 18:492–495.

Zhong B., Xi Z., Goremykin V.V., Fong R., Mclenachan P.A., Novis P.M., Davis C.C., Penny D. 2014. Streptophyte algae and the origin of land plants revisited using heterogeneous models with three new algal chloroplast genomes. Mol. Biol. Evol. 31:177–183.

Zhou M., Li X. 2009. Analysis of synonymous codon-usage patterns in different plant mitochondrial genomes. Mol. Biol. Rep. 36:2039–2046.