MBE Advance Access published October 31, 2013

Streptophyte Algae and the Origin of Land Plants Revisited Using Heterogeneous Models with Three New Algal Chloroplast Genomes

Bojian Zhong,*[1] Zhenxiang Xi,[2] Vadim V. Goremykin,[3] Richard Fong,[1] Patricia A. Mclenachan,[1] Philip M. Novis,[4] Charles C. Davis,[2] and David Penny[1]

[1]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

[2]Department of Organismic and Evolutionary Biology, Harvard University Herbaria

[3]Istituto Agrario San Michele all'Adige Research Center, San Michele all'Adige, Italy

[4]Allan Herbarium, Landcare Research, Lincoln, New Zealand

***Corresponding author:** E-mail: bjzhong@gmail.com.

Associate editor: Charles Delwiche

Abstract

The phylogenetic branching order of the green algal groups that gave rise to land plants remains uncertain despite its fundamental importance to understanding plant evolution. Previous studies have demonstrated that land plants evolved from streptophyte algae, but different lineages of streptophytes have been suggested to be the sister group of land plants. To better understand the evolutionary history of land plants and to determine the potential effects of "long-branch attraction" in phylogenetic reconstruction, we analyzed a chloroplast genome data set including three new chloroplast genomes from streptophyte algae: *Coleochaetae orbicularis* (Coleochaetales), *Nitella hookeri* (Charales), and *Spirogyra communis* (Zygnematales). We further applied a site pattern sorting method together with site- and time-heterogeneous models to investigate the branching order among streptophytes and land plants. Our chloroplast phylogenomic analyses support previous hypotheses based on nuclear data in placing Zygnematales alone, or a clade consisting of Coleochaetales plus Zygnematales, as the closest living relatives of land plants.

Key words: **phylogenomics, chloroplast genomes, land plants, streptophyte algae, heterogeneous models.**

The relationship between green algae and land plants remains uncertain despite its importance to understanding plant evolution. Analyses of both morphological and molecular data have established land plants as a monophyletic group that evolved within streptophyte algae (also known as the charophycean algae). To better understand the colonization of the terrestrial habitat and the evolution of organismal complexity, it is critical to establish which streptophyte groups are most closely related to land plants.

An early study based on four genes from three genomic compartments indicated that Charales were sister to land plants (Karol et al. 2001). In contrast, recent phylogenomic analyses of both chloroplast and nuclear genome data indicated that 1) Coleochaetales alone (Turmel, Gagnon et al. 2009, Turmel, Otis et al. 2009), 2) Zygnematales alone (Turmel et al. 2006; Chang and Graham 2011; Wodniok et al. 2011; Timme et al. 2012; Zhong et al. 2013), or 3) Coleochaetales and Zygnematales combined (Finet et al. 2012; Laurin-Lemay et al. 2012) are sister to land plants. Based on their cytological characteristics, Charales (such as *Chara* and *Nitella*) are large but coenocytic algae with thousands of nuclei per cell (Grant and Borowitzka 1984). In contrast, Coleochaetales (such as *Coleochaete* and *Chaetosphaeridium*) and Zygnematales (such as *Zygnema* and *Spirogyra*) are noncoenocytic organisms that are divided into much smaller cells, each with a single nucleus. In this cytological sense, Coleochaetales or Zygnematales may represent more appropriate sisters to land plants.

Chloroplast genomic data have been proven very useful for helping resolve plant phylogeny (e.g., Jansen et al. 2007; Moore et al. 2007; Zhong et al. 2010; Parks et al. 2012; Wu et al. 2013). In terms of sequenced chloroplast genomes of streptophyte algae, however, there is only one genome currently available for each of the Charales (*Chara vulgaris*) and Coleochaetales (*Chaetosphaeridium globosum*). The paucity of taxon sampling within the most deeply diverged regions of the green plant phylogeny is especially problematic and may lead to long-branch attraction (LBA) artifacts (Hendy and Penny 1989). To ameliorate the problem of LBA, we sequenced three chloroplast genomes from streptophyte algae using next-generation sequencing technology: *Coleochaetae orbicularis* (Coleochaetales), *Nitella hookeri* (Charales), and *Spirogyra communis* (Zygnematales). We then analyzed these data simultaneously in a larger chloroplast genome data set, which includes 72 protein-coding genes (45,879 aligned nucleotide sites) common to 30 land plants and streptophyte algae.

It has been demonstrated that fast-evolving sites represent a challenge for phylogenetic inference because they are likely to experience multiple changes that tend to

Letter

mask informative phylogenetic signals (Delsuc et al. 2005). Recent studies reported that the accuracy of chloroplast phylogenomic analyses could be improved by either removing the most rapidly evolving sites (which is more likely to contain misleading phylogenetic information) or by using site-heterogeneous models (Zhong et al. 2010, 2011; Goremykin et al. 2013). As reported in Zhong et al. (2011) and Goremykin et al. (2013), the "observed variability" (OV)-sorting method (Goremykin et al. 2010) identifies not only the most rapidly evolving sites within a data set but also those sites that have a poor fit to model assumptions. We implemented this method to sort the 45,879 sites in our concatenated matrix from most variable to least variable. We then successively removed the most variable sites in increments of 500. The optimal break point for site removal was determined at site 36,879 (i.e., 9,000 sites were removed from the full matrix), where we identified significant improvement in the two Pearson correlations (see fig. 1).

For the fully concatenated (45,879 aligned sites) and reduced OV-sorted (36,879 aligned sites) matrices, we first used the site-homogeneous GTRGAMMA model with the *a posteriori* partitioning strategy (Xi et al. 2012) to infer our maximum likelihood (ML) phylogeny. The *a posteriori* partitioning strategy partitions data based on the Bayesian searches of the matrix using a mixture model (Pagel and Meade 2004), which has recently been shown to produce better ML trees than the commonly used a priori approaches (e.g., partitioning by gene or by codon position; Xi et al. 2012). Here, our ML analyses strongly support the monophyly of land plants (100 bootstrap percentage [BP]) and strongly place Zygnematales as sister to land plants (97 BP and 96 BP for the fully concatenated and reduced OV-sorted matrices, respectively; fig. 2*A*).

Site-homogeneous models are preferred when a single Markov process of substitution can be applied for all sites and at all times, yet many biological sequences cannot be adequately described using a single replacement matrix. In contrast, site-heterogeneous models introduce different categories by regrouping sites with similar profiles of stationary frequencies and are thus more effective at minimizing LBA artifacts (Lartillot and Philippe 2008; Philippe et al. 2011; Kayal et al. 2013). Therefore, we applied two site-heterogeneous mixture models (Lartillot and Philippe 2004; Pagel and Meade 2004) to infer phylogenetic relationships in a Bayesian framework. Similar to our ML phylogeny using the homogeneous model, both site-heterogeneous models support the relationship (Coleochaetales, [Zygnematales, land plants]) using the fully concatenated matrix with 1.0 posterior probability (PP) (fig. 3*A* and supplementary table S1, Supplementary Material online). However, a slightly different relationship ([Zygnematales, Coleochaetales], land plants) was also strongly supported using this OV-sorted matrix (0.96 PP; fig. 3*B*).

Most current phylogenetic methods (e.g., homogeneous and site-heterogeneous models) assume that base composition is stationary over time. Violation of this model assumption, however, may lead to inaccurate tree reconstruction. Compositional heterogeneity is well known to violate the assumptions of substitution models (Lockhart et al. 1994; Foster 2004; Jermiin et al. 2004). Jobson and Qiu (2011) suggest that compositional shifts of plastid proteins, which could lead to such compositional heterogeneity, might allow

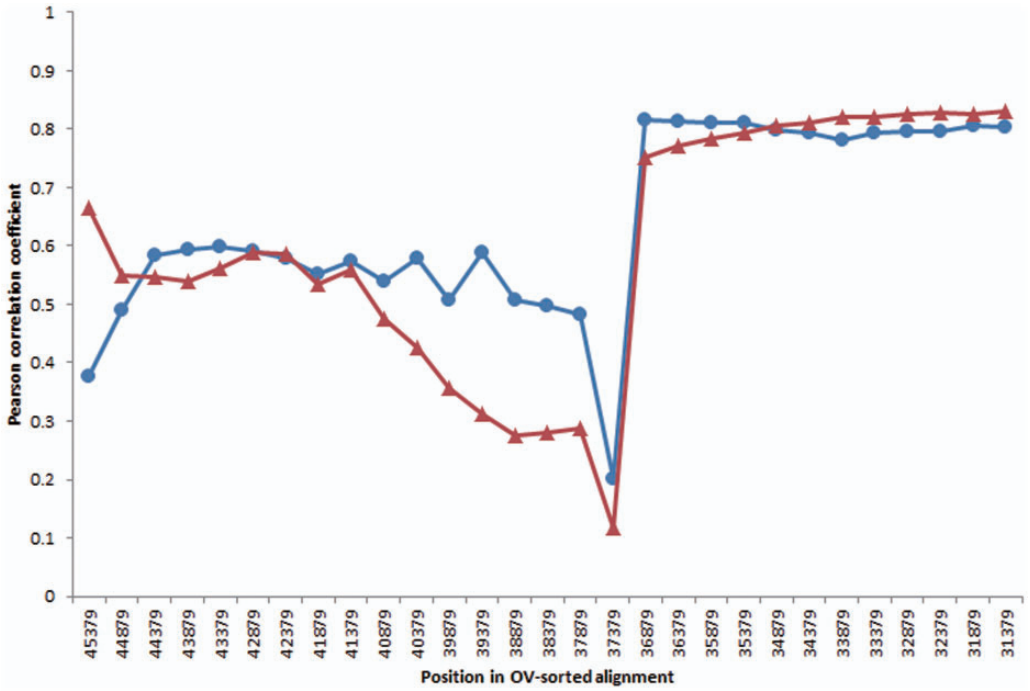

FIG. 1. Pearson correlation results. The line with circles indicates Pearson correlation coefficients (*r*) of ML distances calculated from partitions A (more conserved) and B (less conserved). The line with triangles indicates *r* values of uncorrected *P*-distances and ML distances for B partitions. The *r* values begin to increase dramatically at 36,879 sites remaining.

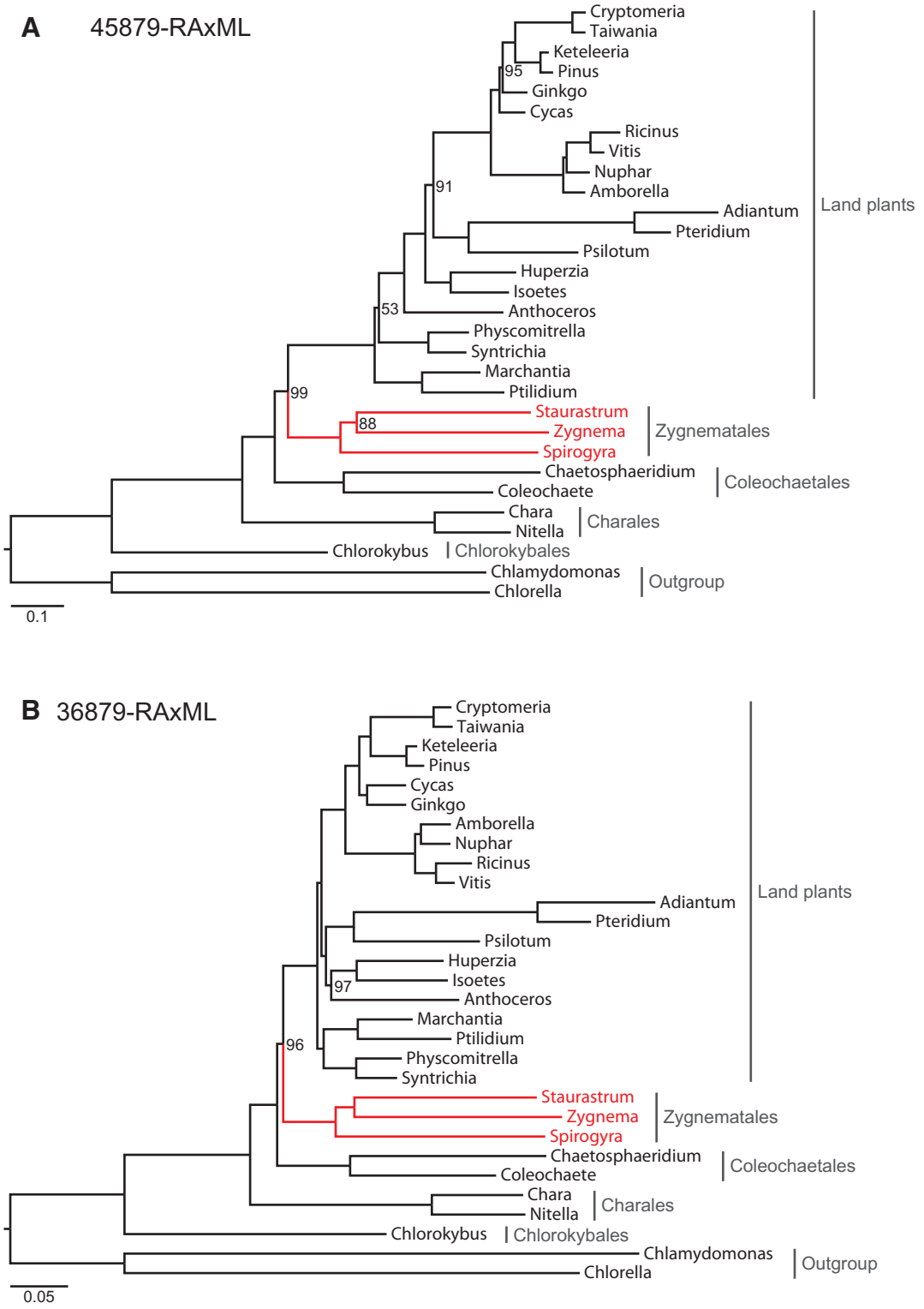

FIG. 2. ML trees using the homogeneous model (GTRGAMMA) with the a posteriori partitioning strategy based on the full (45,879 aligned sites) and reduced OV-sorted (36,879 aligned sites) matrices. Numbers on the tree indicate BP, and nodes with 100 BP are not marked.

streptophyte algae to better deal with environmental stresses on land. To evaluate whether stationarity of composition was violated, we performed posterior predictive tests (Lartillot et al. 2009) for the fully concatenated and reduced OV-sorted matrices. This statistical test indicated that the model assumption of compositional homogeneity is significantly violated in both these matrices (Z-scores are 5.98 and 5.38, respectively; see table 1). Thus, compositional heterogeneity could potentially influence our phylogenetic inference on the origin of land plants. To examine the effect of compositional heterogeneity, we implemented two nonhomogeneous nonstationary (time-heterogeneous) models of DNA sequence evolution in our ML and Bayesian analyses (Galtier and Gouy 1998; Blanquart and Lartillot 2008). When taking

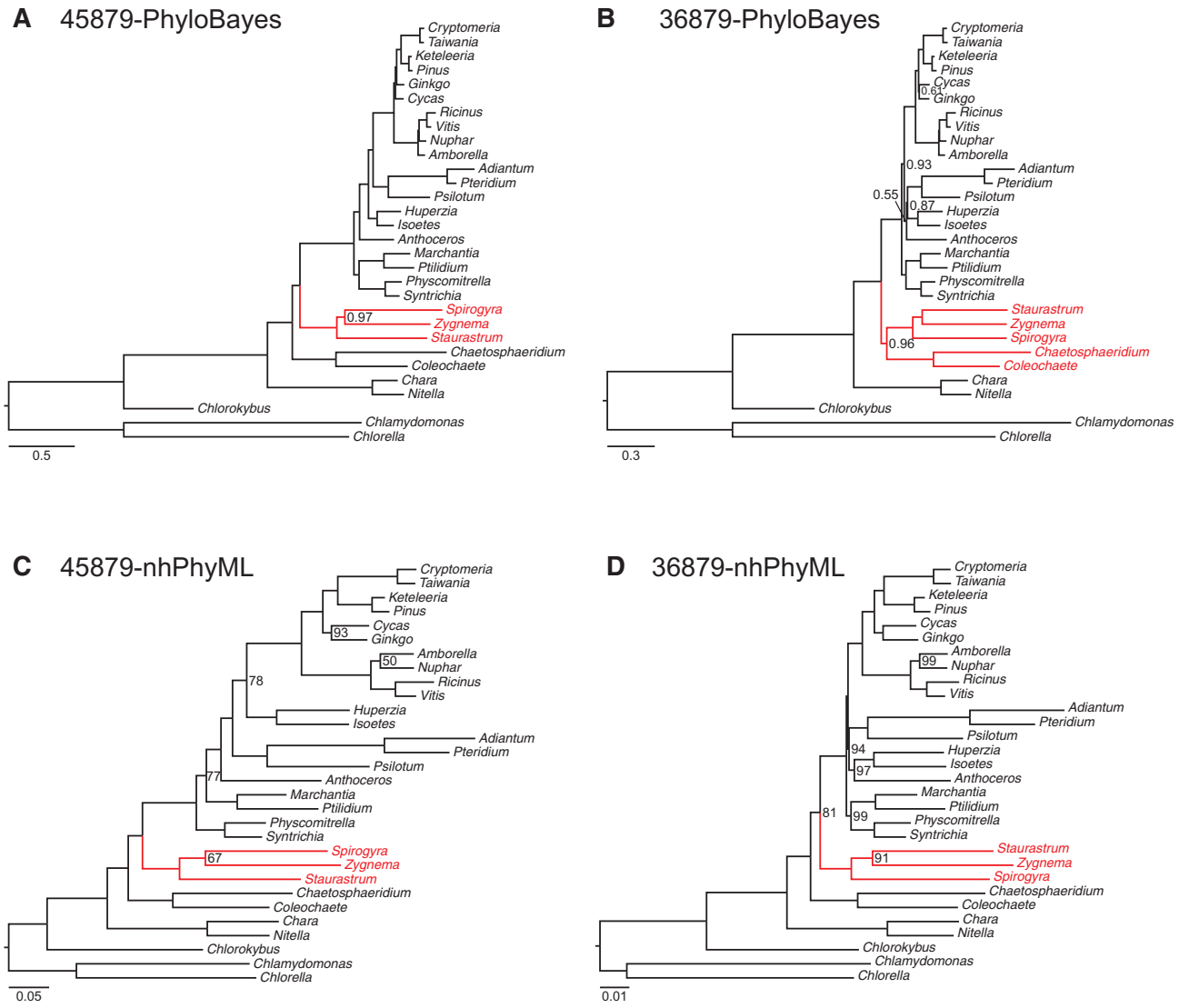

Fig. 3. Phylogenetic trees using the site-heterogeneous model (i.e., the CAT-GTR model) in PhyloBayes and time-heterogeneous model in nhPhyML based on the full (45,879 aligned sites) and OV-sorted (36,879 aligned sites) matrices. Numbers on the tree indicate the Bayesian PP from PhyloBayes and the ML BP from nhPhyML, and nodes with 100 BP or 1.0 PP are not marked.

Table 1. Phylogenetic Analyses Using Bayesian (PhyloBayes) and ML (RAxML and nhPhyML) Estimations.

Data Sets	PhyloBayes	PP	Z-Score	*P* Value	RAxML	BP	nhPhyML	BP
45,879 (full data)	**(CO, [Z, L])**	**1.00**	**5.98**	**0.000**	**(CO, [Z, L])**	**99**	**(CO, [Z, L])**	**100**
38,379	**(CO, [Z, L])**	**0.82**	**5.33**	**0.003**	**(CO, [Z, L])**	**98**	**(CO, [Z, L])**	**96**
37,879	**([Z, CO], L)**	**0.78**	**5.70**	**0.000**	**(CO, [Z, L])**	**73**	**(CO, [Z, L])**	**93**
37,379	**([Z, CO], L)**	**0.91**	**4.20**	**0.007**	**(CO, [Z, L])**	**94**	**(CO, [Z, L])**	**84**
36,879 (OV-sorted data)	**([Z, CO], L)**	**0.96**	**5.38**	**0.000**	**(CO, [Z, L])**	**96**	**(CO, [Z, L])**	**81**
36,379	**([Z, CO], L)**	**0.90**	**5.28**	**0.007**	**([Z, CO], L)**	**100**	**(CO, [Z, L])**	**84**
35,879	**([Z, CO], L)**	**0.98**	**5.15**	**0.003**	**(CO, [Z, L])**	**94**	**(CO, [Z, L])**	**74**
35,379	**([Z, CO], L)**	**0.99**	**4.45**	**0.013**	**([Z, CO], L)**	**52**	**(CO, [Z, L])**	**60**
34,879	**([Z, CO], L)**	**1.00**	**5.85**	**0.000**	**([Z, CO], L)**	**99**	**([Z, CO], L)**	**54**
34,379	**([Z, CO], L)**	**0.99**	**5.30**	**0.000**	**([Z, CO], L)**	**100**	**(CO, [Z, L])**	**62**
33,879	**([Z, CO], L)**	**1.00**	**6.81**	**0.000**	**([Z, CO], L)**	**100**	**(CO, [Z, L])**	**53**
33,379	**([Z, CO], L)**	**1.00**	**6.39**	**0.000**	**([Z, CO], L)**	**100**	**(CO, [Z, L])**	**42**
32,879	**([Z, CO], L)**	**0.96**	**5.76**	**0.000**	**([Z, CO], L)**	**94**	**([Z, CO], L)**	**55**

Note.—CO, Coleochaetales; L, land plants; Z, Zygnematales; PP, Bayesian posterior probability; BP, maximum likelihood bootstrap percentage. The PP and BP values supporting Zygnematales closest to land plants are shown for (CO, [Z, L]) phylogeny, and both values supporting monophyletic relationship of Coleochaetales and Zygnematales are shown for ([Z, CO], L) phylogeny.

into account the compositional heterogeneity using time-heterogeneous models, our results for both the full and OV-sorted matrices consistently supported the relationship (Coleochaetales, [Zygnematales, land plants]) (fig. 3C and D and supplementary table S1, Supplementary Material online).

To evaluate the impact of rapidly evolving sites for estimating branching order among streptophytes and land plants, we further produced 12 shorter alignments by sequentially removing fast-evolving sites in 500 increments using the OV-sorting method, and the number of total sites of these shortened data sets ranges from 38,379 to 32,879 (see table 1 and supplementary table S1, Supplementary Material online). It is striking that the alternative relationship ([Coleochaetales, Zygnematales], land plants) is recovered for 34,879 and 32,879 matrices in all analyses (table1 and supplementary table S1, Supplementary Material online; supplementary figs. S1 and S2, Supplementary Material online). In addition, the relationship (Coleochaetales, [Zygnematales, land plants]) was rejected at $P = 0.05$ for five matrices (i.e., 36,379, 34,879, 34,379, 33,879, and 33,379 aligned sites) using the approximately unbiased (AU) test (Shimodaira 2002) (supplementary table S1, Supplementary Material online). It is noteworthy that none of our analyses here recovered Charales or Coleochaetales alone as the sister group to land plants. Moreover, these two alternative hypotheses were rejected at $P = 0.05$ using the AU test for the 45,879 and 36,879 matrices.

By removing the most rapidly evolving sites and using site- and time-heterogeneous models that reduce both forms of systematic errors (i.e., LBA and compositional heterogeneity), our plastid genomic data indicates that Charales or Coleochaetales alone are not the sister group to land plants. Instead, Zygnematales, or a clade containing Coleochaetales plus Zygnematales, appear to be the closest living relatives of land plants. This result is also in agreement with previous nuclear data analyses (Wodniok et al. 2011; Finet et al. 2012; Laurin-Lemay et al. 2012; Timme et al. 2012; Zhong et al. 2013). One likely explanation for this phylogenetic uncertainty is that Coleochaetales, Zygnematales, and land plants appear to have diverged rapidly during their early evolution (Stebbins and Hill 1980). It is likely important in this context that it was not the "coenocytic" lineage within the Charales that gave rise to land plants. Nevertheless, it is also important to understand the reasons for some green algae (e.g., Charalaes, Zygnematales, and Dasycladales) becoming larger during evolution, and it may be a key for unlocking the origin of land plants.

Materials and Methods

DNA Sequencing and Data Assembly

Nitella hookeri was collected from Lake Wiritoa, Wanganui, New Zealand. *Spirogyra communis* was cultured on BG11 medium (Rippka et al. 1979) from material collected from the Styx River at the Spencerville Road Bridge, Christchurch, New Zealand. *Coleochaete orbicularis* samples were ordered from the Culture Collection of Algae at The University of Texas at Austin (http://web.biosci.utexas.edu/utex, last accessed October 28, 2013) and grown on Modified Bold 3N media. Total genomic DNA (~50 ng) from all three algae was extracted using the Qiagen Plant DNeasy kit according to the manufacturer's protocols and then sequenced using Illumina MiSeq platform with 100-bp paired-end reads. The short reads were filtered with the error probability <0.05 and were then assembled using Velvet (Zerbino and Birney 2008). The contigs were further assembled and compared with complete chloroplast genomes available using Geneious software version 5.6 (www.geneious.com, last accessed October 28, 2013). Protein-coding genes were annotated using DOGMA (Wyman et al. 2004) with manual correction. Each protein-coding gene from 30 taxa was aligned using MUSCLE (Edgar 2004) and trimmed to exclude poorly aligned positions using Gblocks (Castresana 2000) with default settings. These alignments were concatenated to generate a matrix of 45,879 sites.

Removal of Most Rapidly Evolving Sites

The OV-sorting method (Goremykin et al. 2010) was used to rank the full concatenated alignment from the most to least variable sites based on the measurement of observed variability of each alignment position. The most variable sites were then successively removed from the original matrix, in increments of 500. For each step, two data partitions were obtained: 1) "A" partition that consists of all positions except the most variable 500, 1,000, . . . , 9,000 sites and 2) "B" partition that contains the most variable 500, 1,000, . . . , 9,000 sites. After model fitting was applied to each partition using ModelTest (Posada and Crandall 1998), the ML distances for the A and B partitions were calculated, as well as the uncorrected *P*-distances for each "B" partition using PAUP* (Swofford 2002). Two Pearson correlation analyses of pairwise distances were conducted at each step: 1) correlation of the ML distances for A and B partitions and 2) correlation of the ML and uncorrected *P*-distances for B partitions. The stopping point for site removal was determined as the point at which the two correlations showed marked improvement (Goremykin et al. 2010) (see fig. 1).

Phylogenetic Analyses

ML analyses were performed using RAxML (Stamatakis 2006) with the site-homogeneous GTRGAMMA model and the *a posteriori* data partitioning strategy. Two site-heterogeneous Bayesian analyses were then implemented using 1) PhyloBayes (Lartillot et al. 2009) under the CAT-GTR model (Lartillot and Philippe 2004) that accounts for across-site heterogeneities, and 2) BayesPhylogenies (Pagel and Meade 2004) under a "reversible-jump" mixture model (Pagel and Meade 2008) that fits more than one model of sequence evolution to the data. Two independent MCMC analyses were run for 5,000 cycles in PhyloBayes and 10 million generations in BayesPhylogenies. Convergence was checked based on time-series plots of the likelihood scores using Tracer (http://tree.bio.ed.ac.uk/software/tracer/, last accessed October 28, 2013). The posterior predictive test was used to measure compositional heterogeneity in PhyloBayes. Two nonhomogeneous

nonstationary models that account for compositional heterogeneity were applied in both ML and Bayesian analyses, i.e., 1) the nonstationary and nonhomogeneous model of DNA sequence evolution (Galtier and Gouy 1998) as implemented in nhPhyML (Boussau and Gouy 2006) that specifies different GC contents for each lineage in a likelihood framework, and 2) the CAT-BP model (Blanquart and Lartillot 2008) in nhPhyloBayes that considers compositional heterogeneity between lineages by introducing breakpoints along the branches.

The AU test (Shimodaira 2002) was conducted in scaleboot (Shimodaira 2008), with the site log-likelihood scores estimated in RAxML using the *a posteriori* partitioning strategy.

Supplementary Material

Supplementary figures S1, S2, and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the editor and two anonymous reviewers for their helpful comments. Z.X. and C.C.D. were supported by National Science Foundation DEB-1120243. P.M.N. was supported by Core funding for Crown Research Institutes from the Ministry of Business, Innovation and Employment's Science and Innovation Group, New Zealand, and the Brian Mason Scientific and Technical Trust.

References

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol.* 25:842–858.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol.* 55:756–768.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17: 540–552.

Chang Y, Graham SW. 2011. Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. *Am J Bot.* 98:839–849.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.* 6:361–375.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Finet C, Timme RE, Delwiche CF, Marlétaz F. 2012. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr Biol.* 22:1456–1457.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol.* 53: 485–495.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol.* 15: 871–879.

Goremykin VV, Nikoforova SV, Bininda-Emonds OPP. 2010. Automated removal of noisy data in phylogenomic analyses. *J Mol Evol.* 71: 319–331.

Goremykin VV, Nikiforova SV, Biggs PJ, Zhong B, De Lange P, Martin W, Woetzel S, Atherton RA, McLenachan PA, Lockhart PJ. 2013. The evolutionary root of flowering plants. *Syst Biol.* 62:51–62.

Grant BR, Borowitzka MA. 1984. The chloroplasts of giant-celled and coenocytic algae: biochemistry and structure. *Bot Rev.* 50:267–307.

Hendy M, Penny D. 1989. A framework for the quantitative study of evolutionary trees. *Syst Zool.* 38:297–309.

Jansen RK, Cai Z, Raubeson LA, et al. (16 co-authors). 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.

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol.* 53:638–643.

Jobson RW, Qiu Y. 2011. Amino acid compositional shifts during streptophyte transitions to terrestrial habitats. *J Mol Evol.* 72: 204–214.

Karol KG, McCourt RM, Cimino MT, Delwiche CF. 2001. The closest living relatives of land plants. *Science* 294:2351–2353.

Kayal E, Roure B, Philippe H, Collins AG, Lavrov DV. 2013. Cnidarian phylogenetic relationships as revealed by mitogenomics. *BMC Evol Biol.* 13:5.

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

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. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos Trans R Soc Lond B Biol Sci.* 363:1463–1472.

Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. *Curr Biol.* 22:R593–R594.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol.* 11:605–612.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc Natl Acad Sci U S A.* 104:19363–19268.

Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol.* 53:571–581.

Pagel M, Meade A. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos Trans R Soc Lond B Biol Sci.* 363:3955–3964.

Parks M, Cronn RC, Liston A. 2012. Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from Pinus L. (Pinaceae). *BMC Evol Biol.* 12:100.

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

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J Gen Microbiol.* 111:1–61.

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

Shimodaira H. 2008. Testing regions with nonsmooth boundaries via multiscale bootstrap. *J Stat Plan Infer.* 138:1227–1241.

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

Stebbins GL, Hill GJC. 1980. Did multicellular plants invade the land? *Am Nat.* 115:342–353.

Swofford DL 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Timme RE, Bachvaroff TR, Delwiche CF. 2012. Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One* 7:e29696.

Turmel M, Gagnon MC, O'Kelly CJ, Otis C, Lemieux C. 2009. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Mol Biol Evol.* 26:631–648.

Turmel M, Otis C, Lemieux C. 2006. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. *Mol Biol Evol.* 23:1324–1338.

Turmel M, Otis C, Lemieux C. 2009. The chloroplast genomes of the green algae *Pedinomonas minor*, *Parachlorella kessleri*, and *Oocystis solitaria* reveal a shared ancestry between the Pedinomonadales and Chlorellales. *Mol Biol Evol.* 26:2317–2331.

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

Wu CS, Chaw SM, Huang YY. 2013. Chloroplast phylogenomics indicates that *Ginkgo biloba* is sister to cycads. *Genome Biol Evol.* 5: 243–254.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20:3252–3255.

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S, Davis CC. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc Natl Acad Sci U S A.* 109:17519–17524.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–829.

Zhong B, Deusch O, Goremykin VV, Penny D, Biggs PJ, Atherton RA, Nikiforova SV, Lockhart PJ. 2011. Systematic error in seed plant phylogenomics. *Genome Biol Evol.* 3:1340–1348.

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, Yonezawa T, Zhong Y, Hasegawa M. 2010. The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol Biol Evol.* 27:2855–2863.