The mitochondrial phylogeny of land plants shows support for Setaphyta under composition-heterogeneous substitution models

Filipe Sousa [Corresp., 1] , **Peter Civáň** [1, 2, 3] , **João Brazão** [1] , **Peter Foster** [4] , **Cymon Cox** [Corresp. 1]

[1] Centro de Ciências do Mar, Universidade do Algarve, Faro, Portugal

[2] INRA, Université d'Auvergne (Clermont-Ferrand I), Clermont-Ferrand, France

[3] Manchester Institute of Biotechnology, School of Earth and Environmental Sciences, University of Manchester, Machester, United Kingdom

[4] Department of Life Sciences, Natural History Museum, London, United Kingdom

Corresponding Authors: Filipe Sousa, Cymon Cox
Email address: fpsousa@ualg.pt, cymon@ualg.pt

Congruence among analyses of plant genomic data partitions (nuclear, chloroplast and mitochondrial) is a strong indicator of accuracy in plant molecular phylogenetics. Recent analyses of both nuclear and chloroplast genome data of land plants (embryophytes) have, controversially, been shown to support monophyly of both bryophytes (mosses, liverworts, and hornworts) and tracheophytes (lycopods, ferns, and seed plants), with mosses and liverworts forming the clade Setaphyta. However, relationships inferred from mitochondria are incongruent with these results, and typically indicate paraphyly of bryophytes with liverworts alone resolved as the earliest-branching land plant group. Here, we reconstruct the mitochondrial land plant phylogeny from a newly compiled data set. When among-lineage composition heterogeneity is accounted for in analyses of codon-degenerate nucleotide and amino acid data, the clade Setaphyta is recovered with high support, and hornworts are supported as the earliest-branching lineage of land plants. These new mitochondrial analyses demonstrate partial congruence with current hypotheses based on nuclear and chloroplast genome data, and provide further incentive for revision of how plants arose on land.

1 # The mitochondrial phylogeny of land plants shows

2 # support for Setaphyta under composition-

3 # heterogeneous substitution models

4
5

6 Filipe de Sousa[1], Peter Civáň[1,2,3], João Brazão[1], Peter G. Foster[4], Cymon J. Cox[1*]

7
8

9 [1]Centro de Ciências do Mar, Universidade do Algarve, Gambelas, 8005-319 Faro, Portugal
10

11 [2]Manchester Institute of Biotechnology, School of Earth and Environmental Sciences, University
12 of Manchester, Williamson Building, Manchester M13 9PL, United Kingdom

13

14 [3]INRA-Université Clermont-Auvergne, UMR 1095 GDEC, Clermont-Ferrand, France

15

16 [4]Department of Life Sciences, Natural History Museum, London SW7 5BD, United Kingdom

17

18 Corresponding Author: Cymon J. Cox[1]
19 [1]Centro de Ciências do Mar, Universidade do Algarve, Gambelas, 8005-319 Faro, Portugal
20 Email address: cymon.cox@googlemail.com

21
22
23
24

25 ## Abstract

26 Congruence among analyses of plant genomic data partitions (nuclear, chloroplast and

27 mitochondrial) is a strong indicator of accuracy in plant molecular phylogenetics. Recent

28 analyses of both nuclear and chloroplast genome data of land plants (embryophytes) have,

29 controversially, been shown to support monophyly of both bryophytes (mosses, liverworts, and

30 hornworts) and tracheophytes (lycopods, ferns, and seed plants), with mosses and liverworts

31 forming the clade Setaphyta. However, relationships inferred from mitochondria are incongruent

32 with these results, and typically indicate paraphyly of bryophytes with liverworts alone resolved

33 as the earliest-branching land plant group. Here, we reconstruct the mitochondrial land plant

34 phylogeny from a newly compiled data set. When among-lineage composition heterogeneity is

35 accounted for in analyses of codon-degenerate nucleotide and amino acid data, the clade

36   Setaphyta is recovered with high support, and hornworts are supported as the earliest-branching

37   lineage of land plants. These new mitochondrial analyses demonstrate partial congruence with

38   current hypotheses based on nuclear and chloroplast genome data, and provide further incentive

39   for revision of how plants arose on land.

40

41   **Introduction**

42   The embryophytes, or land plants, share a green algal ancestor (McCourt, Delwiche & Karol

43   2004) that colonized terrestrial environments between 515.1- 470.0 Ma (Morris et al. 2018) and

44   comprise gametophyte-dominant lineages, collectively known as bryophytes, and a sporophyte-

45   dominant lineage, the tracheophytes. Establishing the phylogenetic relationships between

46   bryophytes (mosses, liverworts and hornworts) and tracheophytes (a monophyletic lineage that

47   includes lycopods, ferns and seed plants) is therefore fundamental to understanding the evolution

48   of plants on land. However, phylogenetic inferences of relationships among embryophytes drawn

49   from molecular data of the nuclear (Finet et al. 2010; Wodniok et al. 2011; Wickett et al. 2014),

50   chloroplast (Cox et al. 2014; Ruhfel et al. 2014; Zhong et al. 2013; Bell et al. 2020), and

51   mitochondrial (Turmel, Otis & Lemieux 2013; Liu et al. 2014; Bell et al. 2020) genomes have

52   long remained conflicting. These incongruences are likely due to molecular evolutionary

53   processes that are especially apparent at deep timescales, such as multiple substitutions on the

54   same site, that lead to loss of phylogenetic signal, and heterogeneity in substitution process

55   patterns among sites and among lineages (Cox 2018). Phylogenetic patterns commonly observed

56   among embryophytes included a sister-group relationship between hornworts and other

57   embryophytes (Hedderson, Chapman & Rootes 1996; Malek et al. 1996; Nishiyama and Kato

58   1999; Wickett et al. 2014), between liverworts and other embryophytes (Lewis, Mishler &

59   Vilgalys 1997; Karol et al. 2001; Qiu et al. 2006; Gao, Su & Wang 2010; Karol et al. 2010;

60   Clarke, Warnock & Donoghue 2011), or between embryophytes and a clade uniting mosses and

61   liverworts (Karol et al. 2010). An alternative pattern shows an initial split between the bryophytes

62   and the tracheophytes (Hori, Lim & Osawa 1985; Nishiyama et al. 2004; Goremykin and

63   Hellwig 2005; Cox et al. 2014; Wickett et al. 2014; Puttick et al. 2018; Sousa et al. 2019;

64   Leebens-Mack et al. 2019), implying the monophyly of both lineages. Nevertheless, the results

65   are very much dependent on data and methodology, with authors typically presenting competing

66   hypotheses. For instance, several recent phylogenomic analyses based on large nuclear data sets

67   and extensive taxon sampling (e.g. Wickett et al. 2014; Leebens-Mack et al. 2019) have been

68   equivocal. These studies presented monophyletic-bryophyte phylogenies based on multi-species

69   coalescent supertrees, but concatenated analyses of the same data resulted in trees in which the

70   bryophytes were paraphyletic. Consequently, the authors were unable to provide arguments for

71   which hypothesis was to be preferred. Indeed, the efficacy and suitability of using multi-species

72   coalescent summary analyses versus concatenated data analyses for phylogenies with deep

73   timescales is currently a topic of considerable debate (e.g. Tonini et al. 2015; Edwards et al.

74   2016; Springer and Gatesy 2016). However, it should be noted that concatenated analyses of

75   nuclear data do support a monophyletic bryophyte clade when modeling composition

76   heterogeneity across the tree, although restricted analytical conditions, namely reduced taxon and

77   site sampling, are currently necessary to decrease computational complexity when using tree-

78   heterogeneous composition models. For instance, to use these models, Sousa et al (2019)

79   analysed a reduced amino acid data set of 26 taxa and 100 genes, while Puttick et al. (2018)

80   analysed Dayhoff-recoded data, that reduces the amino acid data to only six character states.

81   Nevertheless, these tree-heterogeneous composition models are demonstrably better-fitting and

82   are therefore likely more accurate and reliable than the analyses of larger data sets that used

83   simpler and poorer-fitting models (Cox 2018).

84   Most analyses of land plant relationships have been based on chloroplast data and have

85   typically shown the bryophytes to be paraphyletic (see Table 1 in Cox 2018). More recent

86   phylogenetic analyses using models that account for saturation and composition tree-

87   heterogeneity have, however, indicated that the bryophytes form a monophyletic group, and the

88   authors provided arguments as to why these analyses using better-fitting models are to be

89   preferred (Cox et al. 2014). In contrast, few land plant analyses of mitochondrial data have been

90   presented, but all have indicated that the bryophytes form a paraphyletic group (i.e. Duff and

91   Nickrent 1999; Groth-Malonek et al. 2004; Liu et al. 2014). The most recent and extensive

92   phylogenetic analyses of plant mitochondrial genomes using tree-homogeneous composition

93   models have shown that protein-coding nucleotide data place mosses as the sister-group to the

94   remaining embryophytes, whereas amino acid data show a split between liverworts and the

95   remaining embryophytes (Liu et al. 2014). However, this placement of liverworts as the sister-

96   group to the remaining embryophytes contradicts recent nuclear and chloroplast phylogenies

97   which show high support for the clade Setaphyta, that groups liverworts with mosses (Cox et al.

98   2014; Puttick et al. 2018; Sousa et al. 2019).

99   Our confidence in any evolutionary hypotheses regarding land plant relationships would

100   increase greatly if phylogenetic inferences made from all three plant genomic compartments

101   were not in conflict. In this study we investigate the mitochondrial phylogeny of land plants by

102   applying better-fitting evolutionary models that account for composition tree-heterogeneity to a

103   newly compiled data set of mitochondrial loci that includes sequences from three newly

104   assembled genomes from the Coleochaetales and Zygnematales. We assemble a mitochondrial

105   land plant data set of 26 taxa, which includes all major lineages of land plants and one of the

106   putative most closely-related lineages to land plants, the Zygnematales. Assuming that both

107   bryophytes and tracheophytes are likely monophyletic on the species tree (Sousa et al. 2019), the

108   taxon sampling is deliberately restricted to include a proportional selection of bryophytes (11

109   taxa) and tracheophytes (10) as symmetrical trees improve estimation (Huang and Knowles

110   2009) and minimise long branch attraction (Philippe and Laurent 1998). Where possible, taxa

111   were chosen to span what is currently considered the ancestral node of each bryophyte lineage

112   (or the oldest ancestral node possible given the currently available data), thereby attempting to

113   minimise the length of the sub-tending branch of each bryophyte clade and reduce long-

114   branches. Most importantly, a smaller data set enables us to use better-fitting models that account

115   for among-lineage and among-site composition heterogeneity that are computationally

116   intractable for large data sets. This reduced sampling is in contrast to other studies which

117   typically include many more taxa from highly derived lineages (especially angiosperms) whose

118   inclusion, we maintain, has little impact on the resolution among major lineages (the question

119   under consideration), but severely restricts the complexity of models that can be used and

120   therefore the reliability of the inferences. Notably, a recent large-scale analysis of plant

121   transcriptomes, despite the inclusion of 1155 taxa and 410 genes, was unable to resolve many of

122   the long-standing contentious phylogenetic relationships, such as the relationships among the

123   major lineages of land plants (Fig. 3 in Leebens-Mack et al. 2019). Indeed, rather than just

124   including all available data injudiciously, and thereby restrict the complexity of models that can

125   be applied for phylogenetic inference, it is important to consider which data should be included

126   in a particular analysis: the choice of data should reflect carefully the question being addressed

127   and its suitability for analysis by better-fitting models of molecular evolution.

128

129

Materials & Methods

131

Sampling of mitochondrial genomes

We sampled 21 taxa representing the major lineages of land plants, plus 5 green algae species as outgroup taxa. The taxa sampled for this study were: Coleochaetales (*Chaetosphaeridium globosum, Coleochaete scutata*), Zygnematales (*Closterium baillyanum, Gonatozygon brebissonii, Roya anglica)*, liverworts (*Aneura pinguis. Marchantia polymorpha, Pleurozia purpurea, Treubia lacunosa*), mosses (*Atrichum angustatum, Bartramia pomiformis, Physcomitrella patens, Sphagnum palustre, Tetraphis pellucida*), hornworts (*Megaceros aenigmaticus, Phaeoceros laevis*), lycophytes (*Huperzia squarrosa, Isoetes engelmannii*), pteridophytes (*Ophioglossum californicum, Psilotum nudum*), and spermatophytes (*Brassica napus, Cycas taitungensis, Ginkgo biloba, Liriodendron tulipifera, Oryza sativa, Welwitschia mirabilis*).

Algal cultures for *Coleochaete scutata* (SAG 3.91) and *Gonatozygon brebissonii* (SAG 292) were obtained from the Culture Collection of Algae (SAG) (Georg-August-Universität Göttingen, Germany), and the algal culture of *Roya anglica* (ACOI 799) was obtained from the Coimbra Collection of Algae (ACOI) (Universidade do Coimbra, Coimbra, Portugal). The mitochondrial genomes of the three taxa were sequenced and assembled *de-novo* using standard methods as described in (Civáň et al. 2014), and annotated with the aid of Mitofy (Alverson et al. 2010) and NCBI BLAST analyses (Altschul et al. 1990). The remainder of the genomes were retrieved from NCBI GenBank (Bethesda, USA). The list of species samples, their classification, and the source and accession numbers of the sequences used are shown in Table 1.

152

Alignment and model testing

The sequences of each of 43 mitochondrial protein-coding genes were aligned using the program MAFFT (vers. 6.8; Katoh and Toh 2008). Nucleotide alignments were manually edited in Geneious (vers. 9; http://www.geneious.com) to remove out-of-frame indels, misaligned portions, premature stop codons, positions with less than 50% taxon occupancy, and ambiguities. Genes that were missing with the algal outgroup taxa (Coleochaetales and Zygnematales) or

159   were under 200 bp in length were discarded. The final data set consisted of 36 genes for 26 taxa,

160   with a missing gene occupancy of 9.5%. Amino acid alignments were generated by translation

161   from each of the 36 nucleotide matrices using SeaView (vers. 4.5.4; Gouy, Guindon & Gascuel

162   2009). The best-fitting substitution models for the 36 amino acid alignments were inferred in

163   PartitionFinder (Lanfear et al. 2012) using the BIC selection criterion. The stmtREV (Liu et al.

164   2014) with a gamma-distribution of among-site rates (+G) model was the best-fitting model for

165   all genes but one, for which the chosen model was JTT+G (Jones, Taylor & Thornton 1992).

166

167   *Gene tree estimation and monophyly test*

168

169   Gene trees were estimated from individual nucleotide matrices, using the general time-reversible

170   model (Rodriguez et al. 1990) with a gamma rate distribution and estimated base frequencies

171   (GTR+G+Fest), and from amino acid matrices, using the best-fitting models inferred

172   inPartitionFinder (stmtREV+G+F_{est}, JTT+G+F_{est}). Bayesian MCMC analyses were performed

173   with P4 (Foster 2004) using tree-homogeneous (henceforth referred to as CV1, i.e. one

174   composition vector) and tree-heterogeneous (NDCH) composition models. Each analysis had

175   two independent runs which were assessed for convergence by calculation of the marginal

176   likelihood (chains were considered to have converged if they differed by <10 units) and the

177   average standard-deviation of split support between trees sampled from the posterior

178   distributions (chains were considered to have converged if <0.01). Posterior predictive

179   simulations of the X^2 test statistic of composition homogeneity was used to assess composition

180   fit (Foster 2004). For each gene, 50% majority-rule consensus trees were generated from the

181   best-fitting analyses of the nucleotide and amino acid data.

182   Gene tree topologies inferred from each of the 36 amino acid alignments were tested

183   using "gene genealogy interrogation" (GGI; Arcila et al. 2017) to ascertain whether the non-

184   monophyly of the five major lineages under scrutiny (hornworts, liverworts, mosses,

185   tracheophytes, and the outgroups) was statistically supported by the data. We were to consider

186   any gene that supported the non-monophyly of one of these clades as aberrant and not suitable

187   for inclusion in the combined analyses, as the monophyly of hornworts, liverworts, mosses,

188   tracheophytes and embryophytes has been consistently recovered in molecular phylogenies (e.g.

189   Qiu *et al.* 2006; Liu *et al.* 2014). Optimal trees for each gene were compared to each of fifteen

190   constraint trees representing all possible topologically resolved combinations of the five

191   monophyletic groups, using a nonparametric bootstrapping test. The results were assessed for

192   statistical significance with the Approximately Unbiased (AU) test (Shimodaira 2002) in

193   CONSEL (vers. 0.1k; Shimodaira and Hasegawa 2001). Optimal trees were estimated in

194   RAxML (MPI-compiled vers. 8.2.8; Stamatakis 2014) using the model stmtREV+G+F_{est} and

195   starting from 20 random trees. Constraint trees were written in Newick format with internal

196   branches within each of the five major clades collapsed to a polytomy. Each constraint tree was

197   optimized in RAxML under the stmtREV+G+F_{est} model. Constraint topologies were considered

198   statistically supported by the data if the p-value of the AU test was equal or greater than 0.05,

199   meaning that the monophyly of each of the five lineages could not be rejected. In every gene, at

200   least one of the constraint topologies was supported by the data, meaning that the monophyly of

201   each clade could not be rejected, thus all 36 genes were included in downstream analyses.

202

203   *Analyses of concatenated nucleotide data*

204   A nucleotide alignment with 24,864 characters was obtained from the concatenation of the 36

205   individual genes. A second concatenated nucleotide matrix was constructed by codon-degenerate

206   recoding of the data where ambiguity codes are used to negate synonymous substitutions

207   (Criscuolo and Gribaldo 2010; Cox et al. 2014). Bayesian MCMC analyses were performed on

208   the concatenated and codon-degenerate nucleotide matrices using a tree-homogeneous

209   composition model (CV1; 2 replicates) and the tree-heterogeneous composition (NDCH2; 4

210   replicates) model, as implemented on P4. In contrast to the original NDCH model (Foster 2004)

211   that allows an *a priori* defined number of compositions to evolve on the tree, the NDCH2 model

212   estimates a separate composition for each node of the tree, constrained by a sampled Dirichlet

213   prior on how much the composition vectors may differ from the empirical composition. Model

214   fit to composition heterogeneity was inferred during the Bayesian MCMC with posterior

215   predictive simulations of the X^2 statistic of composition homogeneity, where p-values equal or

216   greater than 0.05 indicate acceptance of the model. The GTR+G+F_{est} model of substitution was

217   used for for all MCMC of nucleotide and degenerate nucleotide data. Marginal likelihoods were

218   estimated in P4 according to the Eqn 16 method of Newton and Raftery (1994).

219

220   *Analyses of concatenated amino acid data*

221 An amino acid alignment with 8,288 characters was obtained by concatenation of the amino acid

222 translations of the 36 genes. Bayesian MCMC analysis was performed on the concatenated

223 amino acid data using both tree-homogeneous composition (stmtREV+G+F_{CV1}; 2 replicates) and

224 tree-heterogeneous composition (stmtREV+G+F_{NDCH2}; 4 replicates) models, with the fit of the

225 composition evaluated by posterior predictive simulations as described for the nucleotide data. In

226 addition, a Bayesian MCMC analysis was also performed using PhyloBayes (MPI-compiled

227 vers. 1.6; Lartillot, Lepage & Blanquart 2009) under the model stmtREV+G+F_{CAT}.

228 Alignments of individual genes, the concatenated data, and the resulting tree files of each

229 analysis are available on Zenodo (doi: 10.5281/zenodo.3554149). All ML analyses of the

230 concatenated nucleotide and amino acid data sets were consistent with the homogeneous

231 Bayesian MCMC analyses and are not reported here, but the resulting tree files are also available

232 on Zenodo.

233

234 **Results**

235 *Nucleotide and codon-degenerate data*

236 All individual genes were best-fit by a tree-heterogeneous composition model with two

237 composition vectors on the tree (CV2; Table S1). Majority-rule consensus trees resulting from

238 the best-fitting Bayesian MCMC analyses of individual genes had low resolution in general, but

239 liverworts were supported (> 95% posterior probability (PP)) as the earliest-branching lineage in

240 two genes (*nad*3 and *nad*5), whereas the mosses were supported as the earliest-branching lineage

241 in one gene (*ccm*C). All other resolutions of the bryophyte lineages relative to the tracheophyte

242 clade were not statistically supported, and the Setaphyta clade was not resolved in any gene tree.

243 Bayesian MCMC analysis of the concatenated nucleotide data set, assuming a

244 homogeneous composition (CV1), resulted in a tree with mosses as the sister-group to the

245 remaining land plants (PP= 1.0; Figure 1a; Figure S1), and hornworts as sister-group to the

246 tracheophytes (PP= 1.0). In contrast, the analysis of the degenerate data set under a

247 homogeneous base composition (Figure 1c; Figure S2) returned a tree with liverworts the

248 earliest-branching lineage of embryophytes (PP= 1.0) and hornworts the sister-group to

249 tracheophytes (PP= 0.93). However, the homogeneous model was rejected for both data sets by

250 the posterior predictive simulation of the X^2 statistics of homogeneity, with a tail-area probability

251 of 0.0, thereby indicating that the data were not composition homogeneous. The tree-

252   heterogeneous composition analysis (NDCH2) of the concatenated nucleotide data also resulted

253   in mosses supported as the earliest-diverging land plant lineage (PP= 0.98; Figure 1b; Figure

254   S3), but placed liverworts as the sister-group to tracheophytes (PP= 0.94). The NDCH2 model

255   was a good fit to the data according to the posterior predictive simulation (X^2 p= 0.99). When

256   analysing the codon-degenerate data, the NDCH2 model recovered hornworts as the sister-group

257   to the remaining land plants with full branch support (PP= 1.0; Figure 1d; Figure S4), and

258   mosses fully supported (PP= 1.0) as the sister-group to liverworts (i.e. the clade Setaphyta). The

259   NDCH2 model was not a good statistical fit to the data for the best scoring MCMC run,

260   according to the X^2 posterior predictive test (X^2 p= 0.038; Figure S4).

261

262   *Amino acid data*

263   Individual mitochondrial protein alignments were best-fit by both homogeneous and

264   heterogeneous composition models, with some being best-fit by a model with up to four

265   compositions (CV4), indicating that they are highly heterogeneous in composition among

266   lineages (Table S1). Majority-rule consensus trees of best-fitting Bayesian MCMC analyses of

267   individual proteins were poorly supported regarding relationships among bryophyte lineages.

268   Indeed, only one tree (*ccm*C) showed any statistically supported resolution and placed the

269   mosses as the earliest-branching lineage of embyrophytes. Although not statistically supported,

270   one amino-acid tree showed bryophytes as a monophyletic group (*rps*7), and the clade Setaphyta

271   was present in three others (*atp*4, *rpl*2, and *sdh*4).

272   When the concatenated amino acid data were analysed with a tree-homogeneous

273   composition model (CV1), the resulting tree recovered liverworts as the sister-group to the

274   remaining embryophytes without statistically significant support (PP= 0.82), and hornworts as

275   the sister-group to tracheophytes, also without support (PP= 0.82; Figure 1e; Figure S5). The

276   homogeneous composition model did not fit the data (X^2 p= 0.0). The Phylobayes

277   stmtREV+G+F_{CAT} analysis (stationary, data-heterogeneous composition model) of the

278   concatenated amino acid data resulted in an unsupported placement of liverworts as the sister-

279   group to the remaining embryophytes (PP= 0.89; Figure S6). A posterior predictive composition

280   homogeneity test using Phylobayes (*readpb_mpi* parameter -*comp*) showed that the data rejected

281   the model and that the data were therefore composition tree-heterogeneous (p= 0.0). When the

282   tree-heterogeneous composition NDCH2 model was used to model the concatenated protein data,

283   the separate analyses did not converge on the same tree topology, although the NDCH2 model

284   was a good fit (X^2 p= 0.1022). The tree obtained from the analyses with the best marginal

285   likelihood (-lnLh 142829.1129) supported hornworts as the earliest-branching lineage (PP= 1.0),

286   with the liverworts as the sister-group to mosses (PP=1.0; Figs.1f, 2, S7).

287

288   Discussion

289

290   *Gene tree discordance in mitochondrial data is likely due to mis-modeling and insufficient*

291   *phylogenetic signal*

292   Alternative hypothesis testing using nonparametric and parametric bootstrapping has been

293   applied before to the mitochondrial land plant phylogeny to test the fit of the data (Liu et al.

294   2014). Here we chose a different approach, and used the optimized likelihood of constraint trees

295   to identify genes that did not support the monophyly of the major embryophyte lineages

296   (hornworts, liverworts, mosses, and tracheophytes) and of the outgroup. The optimal trees of the

297   36 mitochondrial genes, inferred under maximum-likelihood, show varied topologies, among

298   which none is predominant. Because the four major land plant lineages are known to be

299   monophyletic (as shown by many studies, e.g. Wickett et al. 2014) our concern was to identify

300   gene trees that showed non-monophyly of one of these groups. The strategy we adopted allowed

301   us to discern whether such topologies truly reflect underlying data or whether they are one

302   among different topologies supported by the data. The AU test indicated, in all genes, that the

303   monophyly of each land plant lineage was not statistically contradicted. This result suggests that

304   the observed phylogenetic conflict among mitochondrial gene trees is unlikely to be explained by

305   biological processes, such as horizontal gene transfer or duplication-loss, affecting specific

306   lineages within each of the four major groups, and that any observed paraphyly of major groups

307   on gene trees is probably the result of inadequate data modeling or paucity of phylogenetic

308   signal.

309

310   *Synonymous substitutions are responsible for the placement of mosses as the earliest-branching*

311   *lineage of embryophytes*

312   The tree inferred from the concatenated nucleotide data set of 36 mitochondrial genes shows

313   mosses as the sister-group to the remaining land plants, as previous analyses of mitochondrial

314   nucleotide data have shown (Liu et al. 2014). However, the mosses are replaced by the liverworts

315   in the same position when analysing codon-degenerate recoded data. Codon degenerate recoding

316   is used to eliminate synonymous substitutions, which are unconstrained by selection at the

317   protein level and therefore can be subject to high rates of substitution, and ultimately saturation

318   and loss of phylogenetic signal. Indeed, as the exclusion of synonymous substitutions is

319   sufficient to eliminate phylogenetic signal that supports mosses as the sister-group to the

320   remaining land plants, these results illustrate that despite being the slowest evolving genomic

321   compartment in plants, phylogenetic inferences from highly divergent mitochondrial genomes

322   are also affected by substitutional saturation due to the effect of high substitution rates.

323   Moreover, this observation implies that caution should be taken when invoking biological

324   explanations (e.g. hybridisation, incomplete lineage sorting) to explain incongruence among

325   phylogenies when inadequate modeling of the substitution process may represent a simpler

326   explanation.

327

328   *Codon degenerate nucleotide data and protein data support Setaphyta under tree-heterogeneous*

329   *models of substitution*

330   With the nucleotide data, both the tree-homogeneous and NDCH2 tree-heterogeneous models

331   support mosses as the earliest-diverging group in the embryophytes. The likely incorrect

332   placement of mosses as the earliest-diverging group that is recovered with the best-fitting tree-

333   heterogeneous NDCH2 model suggests that homoplasy driven by high nucleotide substitution

334   rates (saturation) may overwhelm the ability of the model to correct for composition bias, hence

335   the need to use codon-degenerate recoded data in combination with tree-heterogeneous models.

336   Indeed, when codon-degenerate recoded data are used, contrasting supported relationships are

337   obtained under tree-homogeneous and tree-heterogeneous composition models. Whereas using a

338   homogeneous model for the analysis of the codon-degenerate data shows liverworts well

339   supported as the sister-group to other embryophytes, the tree-heterogeneous analysis (NDCH2)

340   model places liverworts as the sister-group to the mosses (clade Setaphyta), with maximum

341   support, and hornworts as the sister-group to all other embryophytes, also with maximum branch

342   support. These results demonstrate that the phylogenetic signal contained in non-synonymous

343   sites is also subject to composition biases and that tree-heterogeneous composition models are

344   required to model the data effectively. Contrasting results were also obtained when the amino

345   acid data were analysed with tree-homogeneous and tree-heterogeneous models. The tree

346   inferred under the poorly-fitting (P=0.0) homogeneous model (CV1) resolves liverworts as the

347   sister-group to the remaining land plants (PP=0.82). By contrast, the Bayesian MCMC run with

348   the highest marginal likelihood under the NDCH2 (fitting) model shows strong support for the

349   Setaphyta clade (PP= 1.0 – mosses plus liverworts) with the hornworts as the earliest-branching

350   lineage of embryophytes (PP= 1.0). Liu et al. (2014) observed topological congruence between

351   mitochondrial nucleotide and protein data that placed liverworts as the sister-group to all other

352   embryophytes, but this placement of liverworts received low branch support in different

353   analyses, and thus no firm conclusions regarding these cladogenic events were put foward. In

354   contrast, our analyses of the codon-degenerate nucleotide data and amino acid data using a

355   better-fitting tree-heterogeneous model resulted in well supported, congruent topologies,

356   strengthening the argument in favour of the analyses presented here, that show support for the

357   clade Setaphyta.

358

359   *The land plant mitochondrial phylogeny is partially congruent with nuclear and chloroplast*

360   *phylogenies*

361   In contrast to previous analyses of the land plant mitochondrial phylogeny, we show that both

362   nucleotide and amino acid data carry signal that joins mosses and liverworts as sister lineages

363   (clade Setaphyta). This phylogenetic signal is typically obscured due to substitution saturation in

364   the nucleotide data and among-lineage composition bias in both the nucleotide and amino acid

365   data. In the nucleotide data, phylogenetic signal supporting mosses as the sister-group to the

366   remaining land plants is eliminated when codon-degenerate recoded data is analysed, and instead

367   the liverworts are found as the sister-group to all the remaining embryophytes under tree-

368   homogeneous composition models. However, it is only when a combination of codon-degenerate

369   recoding and a better-fitting tree-heterogeneous composition model is used that the mosses and

370   liverworts appear resolved as sister taxa, therefore suggesting that both substitutional saturation

371   and among-lineage composition heterogeneity are important evolutionary processes to be

372   modeled in the nucleotide data. Similarly, the unsupported placement of liverworts as the

373   earliest-branching lineage is obtained using tree-homogeneous composition models with the

374   amino acid data, but better-fitting tree-heterogeneous composition models again support the

375   mosses plus liverwort clade.

376   Support for the moss-liverwort sister-group relationship has been found in trees

377   previously inferred from nuclear and chloroplast protein-coding data (e.g. Nishiyama et al. 2004;

378   Cox et al. 2014; Puttick et al. 2018; Sousa et al. 2019). The clade can be resolved by

379   mitochondrial data with our analyses, and therefore avoids the necessity of calling upon

380   biological explanations, such as hybridisation, to account for incongruence among the

381   phylogenies of the three plant genomes regarding the placement of mosses and liverworts.

382   However, if the placement of the hornworts as the earliest-branching lineage of embryophytes

383   does indeed reflect the true mitochondrial topology, then it is in conflict with the nuclear and

384   chloroplast data which suggest the bryophytes are likely monophyletic. A biological process

385   involving a rapid divergence of the hornworts from other bryophytes, after the tracheophyte-

386   bryophyte split, and the retention of a copy of the mitochondrion that was lost in all other

387   embryophyte lineages, could be invoked to explain the observed phylogenetic conflict. However,

388   the incongruence of the mitochondrial data could, of course, still be a result of mis-modeling or

389   lack of phylogenetic signal. It is likely that further sampling of mitochondrial genomes from

390   hornworts and other bryophyte lineages may aid resolution of the phylogeny, but such analyses

391   would only be informative if they were in combination with the heterogeneous composition

392   models that have been shown here to be necessary to correctly model the underlying processes of

393   mitochondrial evolution.

394

395

396   **Conclusions**

397

398   The main contribution of this study is the demonstration that liverworts are not the sister-group

399   to embryophytes in the land plant mitochondrial phylogeny, unlike earlier analyses of

400   mitochondrial genomes suggested (Liu et al. 2014). Instead, strong support is found for the clade

401   Setaphyta, corroborating support for this clade found in nuclear and plastid genomes, and

402   showing that the mitochondrial phylogeny of land plants is not strongly incongruent with the

403   nuclear and plastid phylogenies. Although the best-scoring tree found by analyses of amino acid

404   data places hornworts as sister-group to embryophytes, the monophyly of bryophytes, which is

405   supported by evidence from nuclear and plastid genomes, cannot be strongly rejected.

406   Importantly, this study also shows that modeling of composition tree-heterogeneity in amino acid

407 data must not be disregarded, even in slower-evolving genomic regions such as plant

408 mitochondria.

409

410

411

412 **References**

413

414 Alverson AJ, Wei XX, Rice DW, Stern DB, Barry K, Palmer JD. 2010. Insights into the

415 evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and

416 *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution* 27: 1436–1448.

417 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.

418 *Journal of Molecular Biology* 215: 403–410.

419 Arcila D, Ortí G, Vari R, Armbruster JW, Stiassny ML, Ko KD, Sabaj MH, Lundberg J, Revell

420 LJ, Betancur-R R. 2017. Genome-wide interrogation advances resolution of recalcitrant

421 groups in the tree of life. *Nature Ecology & Evolution* 1: 0020.

422

423 Bell, D., Lin, Q., Gerelle, W. K., Joya, S., Chang, Y., Taylor, Z. N.,Rothfels, C.J., Larsson, A.,

424 Villarreal, J.C., Li, F.-W., Pokorny, L., Szövényi, P., Crandall-Stotler, B., DeGironimo, L.,

425 Floyd, S.K., Beerling, D.J., Deyholos, M.K., von Konrat, M., Ellis, S., Shaw, A.J., Chen,

426 T., Wong, G.K.-S., Stevenson, D.W., Palmer, J.D., Graham, S.W. (2020). Organellomic

427 data sets confirm a cryptic consensus on (unrooted) land-plant relationships and provide

428 new insights into bryophyte molecular evolution. *American Journal of Botany* 107: 91-115.

429 Civáň P, Foster PG, Embley MT, Seneca A, Cox CJ. 2014. Analyses of charophyte chloroplast

430 genomes help characterize the ancestral chloroplast genome of land plants. *Genome

431 Biology and Evolution* 6: 897–911.

432 Clarke JT, Warnock RCM, Donoghue PCJ. 2011. Establishing a time-scale for plant evolution.

433 *New Phytologist* 192: 266–301.

434 Cox CJ, Li B, Foster PG, Embley TM, Civáň P. 2014. Conflicting phylogenies for early land

435 plants are caused by composition biases among synonymous substitutions. *Systematic

436 Biology* 63: 272–279.

437   Cox CJ. 2018. Land plant molecular phylogenetics: a review with comments on evaluating

438   incongruence among phylogenies. *Critical Reviews in Plant Sciences* 37: 113–127.

439   Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new

440   software for selection of phylogenetic informative regions from multiple sequence

441   alignments. *BMC Evolutionary Biology* 10: 210.

442   Duff RJ, Nickrent DL. 1999. Phylogenetic relationships of land plants using mitochondrial small

443   subunit rDNA sequences. *American Journal of Botany* 86: 372-386.

444   Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, Zhong B, Wu S, Lemmon

445   EM, Lemmon AR, Leaché AD. 2016. Implementing and testing the multispecies

446   coalescent model: a valuable paradigm for phylogenomics. *Molecular Phylogenetics and*

447   *Evolution* 94: 447–462.

448   Finet C, Timme RE, Delwiche CF, Marlétaz F. 2010. Multigene phylogeny of the green lineage

449   reveals the origin and diversification of land plants. *Current Biology* 20: 2217–2222.

450   Foster PG. 2004. Modeling compositional heterogeneity. *Systematic Biology* 53: 485–495.

451   Gao L, Su YJ, Wang T. 2010. Plastid genome sequencing, comparative genomics, and

452   phylogenomics: Current status and prospects. *Journal of Systematics and Evolution* 48:

453   77–93.

454   Goremykin VV, Hellwig FH. 2005. Evidence for the most basal split in land plants dividing

455   bryophyte and tracheophyte lineages. *Plant Systematics and Evolution* 254: 93-103.

456   Gouy M, Guindon S, Gascuel O. 2009. SeaView version 4: a multiplatform graphical user

457   interface for sequence alignment and phylogenetic tree building. *Molecular Biology and*

458   *Evolution* 27: 221–224.

459   Groth-Malonek M, Pruchner D, Grewe F, Knoop V. 2004. Ancestors of trans-splicing

460   mitochondrial introns support serial sister group relationships of hornworts and mosses

461   with vascular plants. *Molecular Biology and Evolution* 22: 117–125.

462   Hedderson TA, Chapman RL, Rootes, WL. 1996. Phylogenetic relationships of bryophytes

463   inferred from nuclear-encoded rRNA gene sequences. *Plant Systematics and Evolution*

464   *200*: 213-224.

465   Hori H, Lim B-L, Osawa S. 1985. Evolution of green plants as deduced from 5S rRNA

466   sequences. *Proceedings of the National Academy of Sciences U. S. A.* 82: 820–823.

467 Huang H, Knowles LL. 2009. What is the danger of the anomaly zone for empirical
468   phylogenetics? *Systematic Biology* 58: 527–536.
469 Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from
470   protein sequences. *Bioinformatics* 8: 275-282.
471 Karol KG, McCourt RM, Cimino MT, Delwiche CF. 2001. The closest living relatives of land
472   plants. *Science* 294: 2351–2353.
473 Karol KG, Karol KG, Arumuganathan K, Boore JL, Duffy AM, Everett KD, Hall JD, Hansen
474   SK, Kuehl JV, Mandoli DF, Mishler BD, Olmstead RG. 2010. Complete plastome
475   sequences of *Equisetum arvense* and *Isoetes flaccida*: implications for phylogeny and
476   plastid genome evolution of early land plant lineages. *BMC Evolutionary Biology* 10: 321.
477 Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment
478   program. *Brief. Bioinformatics* 9: 286–298.
479 Lanfear R, Calcott B, Ho SY, Guindon S. 2012. PartitionFinder: combined selection of
480   partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology*
481   *and Evolution* 29: 1695–1701.
482 Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for
483   phylogenetic reconstruction and molecular dating. *Bioinformatics* 25: 2286–2288.
484 Leebens-Mack J, Barker M, Carpenter E, Deyholos M, Gitzendanner M, Graham S, Grosse I, Li
485   Z, Melkonian M, Mirarab S, Porsch M, Quint M, Rensing S, Soltis D, Soltis P, Stevenson
486   D, Ullrich K, Wickett N, DeGironimo L, Edger P, Jordon-Thaden I, Joya S, Liu T,
487   Melkonian B, Miles N, Pokorny L, Quigley C, Thomas P, Villarreal J, Augustin M, Barrett
488   M, Baucom R, Beerling D, Benstein R, Biffin E, Brockington S, Burge D, Burris J, Burris
489   K, Burtet-Sarramegna V, Caicedo A, Cannon S, Cebi Z, Chang Y, Chater C, Cheeseman
490   J,Chen T, Clarke N, Clayton H, Covshoff S, Crandall-Stotler B, Cross H, DePamphilis C,
491   Der J, Determann R, Dickson R, Di Stilio V, Ellis S, Fast E, Feja N, Field K, Filatov D,
492   Finnegan P, Floyd S, Fogliani B, Garcia N, Gateble G, Godden G, Goh Q.-Y, Greiner S,
493   Harkess A,Heaney J, Helliwell K, Heyduk K, Hibberd J, Hodel R, Hollingsworth P,
494   Johnson M, Jost R,Joyce B, Kapralov M, Kazamia E, Kellogg E, Koch M, Von Konrat M,
495   Konyves K, Kutchan T, Lam V, Larsson A, Leitch A, Lentz R, Li F.-W, Lowe A, Ludwig
496   M, Manos P, Mavrodiev E, McCormick M, McKain M, McLellan T, McNeal J, Miller R,
497   Nelson M, Peng Y, Ralph P,Real D, Riggins C, Ruhsam M, Sage R, Sakai A, Scascitella

498   M, Schilling E, Schlosser E.-M, Sederoff H, Servick S, Sessa E, Shaw A, Shaw S, Sigel E,
499   Skema C, Smith A, Smithson A, Stewart C, Stinchcombe J, Szovenyi P, Tate J, Tiebel H,
500   Trapnell D, Villegente M, Wang C.-N, Weller S, Wenzel M, Weststrand S, Westwood J,
501   Whigham D, Wu S, Wulff A, Yang Y, Zhu D, Zhuang C, Zuidof J, Chase M, Pires J,
502   Rothfels C, Yu J, Chen C, Chen L, Cheng S, Li J, Li R, Li X, Lu H, Ou Y, Sun X, Tan X,
503   Tang J, Tian Z, Wang F, Wang J, Wei X, Xu X, Yan Z, Yang F, Zhong X, Zhou F, Zhu Y,
504   Zhang Y, Ayyampalayam S, Barkman T, Nguyen N.-P, Matasci N, Nelson D, Sayyari E,
505   Wafula E, Walls R, Warnow T, An H, Arrigo N, Baniaga A, Galuska S, Jorgensen S,
506   Kidder T, Kong H, Lu-Irving P, Marx H, Qi X, Reardon C, Sutherland B, Tiley G, Welles
507   S, Yu R, Zhan S, Gramzow L, Theissen G, Wong G. K.-S. 2019. One thousand plant
508   transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.

509   Lewis LA, Mishler BD, Vilgalys R. 1997. Phylogenetic relationships of the liverworts
510   (Hepaticae), a basal embryophyte lineage, inferred from nucleotide sequence data of the
511   chloroplast gene rbcL. *Molecular Phylogenetics and Evolution* 7: 377–393

512   Liu Y, Cox CJ, Wang W, Goffinet B. 2014. Mitochondrial phylogenomics of early land plants:
513   Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias.
514   *Systematic Biology* 63: 862–878.

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

517   McCourt RM, Delwiche CF, Karol KG. 2004. Charophyte algae and land plant origins. *Trends in*
518   *Ecology & Evolution* 19: 661–666.

519   Morris JL, Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH,
520   Yang Z, Schneider H, Donoghue PC. 2018. The timescale of early land plant evolution.
521   *Proceedings of the National Academy of Sciences U. S. A.* 115: E2274–E2283.

522   Newton MA, Raftery AE. 1994. Approximate Bayesian inference with the weighted likelihood
523   bootstrap. *Journal of the Royal Statistical Society B* 56: 3–48.

524   Nishiyama T, Nishiyama T, Wolf PG, Kugita M, Sinclair RB, Sugita M, Sugiura C, Wakasugi T,
525   Yamada K, Yoshinaga K, Yamaguchi K, Ueda K. 2004. Chloroplast phylogeny indicates
526   that bryophytes are monophyletic. *Molecular Biology and Evolution* 21: 1813–1819.

527   Nishiyama T, Kato M. 1999. Molecular phylogenetic analysis among bryophytes and
528   tracheophytes based on combined data of plasmid coded genes and the 18S rRNA gene.
529   *Molecular Biology and Evolution* 16: 1027–1036.

530   Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? *Current Opinion in*
531   *Genetics and Development* 8: 616–623.

532   Puttick MN, Morris JL, Williams TA, Cox CJ, Edwards D, Kenrick P, Pressel S, Wellman CH,
533   Schneider H, Pisani D, Donoghue PC. 2018. The interrelationships of land plants and the
534   nature of the ancestral embryophyte. *Current Biology* 28: 733–745.

535   Qiu Y-L, Li L, Wang B, Chen Z, Knoop V, Groth-Malonek M, Dombrovska O, Lee J, Kent L,
536   Rest J, Estabrook GF. 2006. The deepest divergences in land plants inferred from
537   phylogenomic evidence. *Proceedings of the National Academy of Sciences U. S. A.* 103:
538   15511–15516.

539   Rodriguez F, Oliver JL, Marin A, Medina JR. 1990. The general stochastic model of nucleotide
540   substitution. *Journal of Theoretical Biology* 142: 485-501.

541   Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to
542   angiosperms--inferring the phylogeny of green plants (Viridiplantae) from 360 plastid
543   genomes. *BMC Evolutionary Biology* 14: 23.

544   Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic*
545   *Biology* 51: 492–508.

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

548   Sousa F, Foster PG, Donoghue PC, Schneider H, Cox CJ. 2019. Nuclear protein phylogenies
549   support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New*
550   *Phytologist* 222: 565–575.

551   Springer MS, Gatesy J. 2016. The gene tree delusion. *Molecular Phylogenetics and Evolution*
552   94, 1-33.

553   Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
554   large phylogenies. *Bioinformatics* 30: 1312-1313.

555   Tonini J, Moore A, Stern D, Shcheglovitova M, Ortí G. 2015. Concatenation and species tree
556   methods exhibit statistically indistinguishable accuracy under a range of simulated
557   conditions. *PLoS currents,* 7.

558   Turmel M, Otis C, Lemieux C. 2013. Tracing the evolution of streptophyte algae and their

559   mitochondrial genome. *Genome Biology and Evolution* 5: 1817–1835.

560   Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter C, Matasci N, Ayyampalayam S,

561   Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR, Wafula E, Der JP, Graham SW,

562   Mathews S, Melkonian M, Soltis DE, Soltis PS, Miles NW, Rothfels C-J, Pokorny L,

563   Shaw AJ, DeGironimo L, Stevenson DW, Surek B, Villarreal JC, Roure B, Philippe H,

564   dePamphilis CW, Chen T, Deyholos MK, Baucom RS, Kutchan TM, Augustin MM, Wang

565   J, Zhang Y, Tian Z, Yan Z, Wu X, Sun X, Wong GK-S, Leebens-Mack J. 2014.

566   Phylotranscriptomic analysis of the origin and early diversification of land plants.

567   *Proceedings of the National Academy of Sciences U. S. A.* 111: E4859–E4868.

568   Wodniok S, Brinkmann H, Glöckner G, Heidel AJ, Philippe H, Melkonian M. Becker B. 2011.

569   Origin of land plants: do conjugating green algae hold the key? *BMC Evolutionary Biology*

570   11: 104.

571   Zhong B, Xi Z, Goremykin VV, Fong R, Mclenachan PA, Novis PM, Davis CC, Penny D. 2013.

572   Streptophyte algae and the origin of land plants revisited using heterogeneous models with

573   three new algal chloroplast genomes. *Molecular Biology and Evolution* 31: 177-183.

574

Table 1(on next page)

Accession table of the 26 samples used in this study.

For each species, the corresponding taxonomic group and NCBI GenBank accession numbers are shown. Accessions marked with an asterisk (*) correspond to newly assembled genomes.

1

species	taxonomic group	Accession
Closterium baillyanum	Zygnematales	NC_022860
Gonatozygon brebissonii	Zygnematales	MK720950 *
Roya anglica	Zygnematales	MK720948 *
Chaetosphaeridium globosum	Coleochaetales	NC_004118
Coleochaete scutata	Coleochaetales	MK720949 *
Megaceros aenigmaticus	hornworts	NC_012651
Phaeoceros laevis	hornworts	NC_013765
Aneura pinguis	liverworts	NC_026901
Marchantia polymorpha	liverworts	NC_001660
Pleurozia purpurea	liverworts	NC_013444
Treubia lacunosa	liverworts	NC_016122
Atrichum angustatum	mosses	NC_024520
Bartramia pomiformis	mosses	NC_024519
Physcomitrella patens	mosses	NC_007945
Sphagnum palustre	mosses	NC_024521
Tetraphis pellucida	mosses	NC_024290
Welwitschia_mirabilis	seed plants	NC_029130
Brassica napus	seed plants	NC_008285
Liriodendron tulipifera	seed plants	NC_021152
Oryza_sativa	seed plants	NC_011033
Cycas taitungensis	seed plants	NC_010303
Ginkgo biloba	seed plants	NC_027976

Ophioglossum californicum	ferns	NC_030900
Psilotum nudum	ferns	NC_030952, KX171639
Huperzia squarrosa	lycophytes	NC_017755
Isoetes engelmannii	lycophytes	FJ010859, FJ176330, FJ390841, FJ536259, FJ628360

2

3

Figure 1

A schematic representation of the topologies obtained from tree-homogeneous and tree-heterogeneous analyses of nucleotide, codon-degenerate nucleotide, and amino acid translation data.

Analyses of nucleotide data place mosses as the earliest-branching lineage of the embryophytes (1a, 1b). Analyses of codon-degenerate and amino acid data under tree-homogeneous models place liverworts as the sister-group to the remaining embryophytes (1c, 1e), whereas analyses under tree-heterogeneous models show support for the clade Setaphyta (1d, 1f).
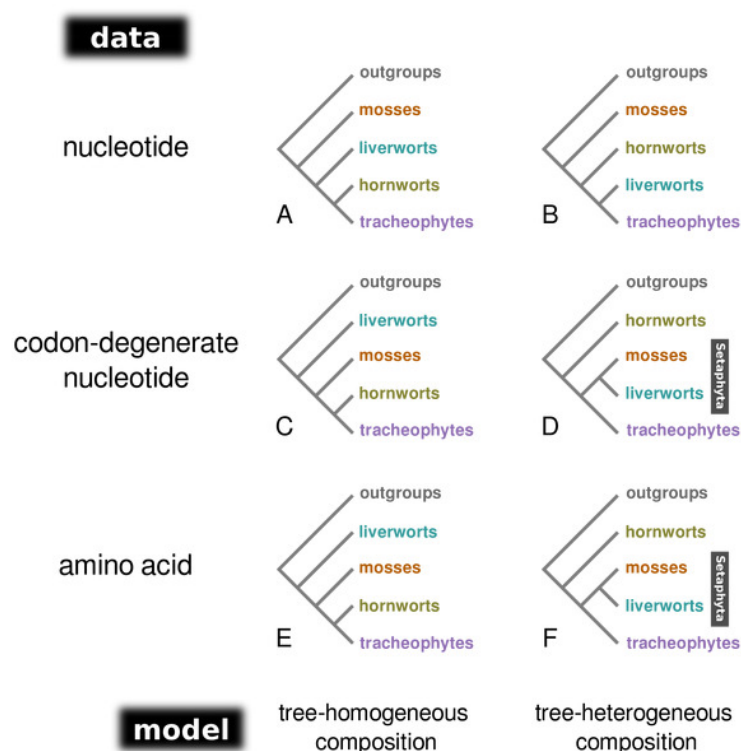
Figure 2

Majority-rule consensus tree inferred from the 36 gene, 26 taxon concatenated amino acid data.

Bayesian MCMC with a tree-heterogeneous composition model NDCH2, marginal likelihood - L_h= 142829.112. Additional analysis statistics can be found in the legend of Fig. S7. All branches fully supported (PP=1.0). Taxa indicated as follows: outgroups, grey; hornworts, olive green; liverworts, cyan blue; mosses, orange; vascular plants, violet.