



11           ABSTRACT. With the rapid reduction in sequencing costs of high-throughput genomic  
12 data, it has become commonplace to use hundreds of genes/sites to infer phylogeny of any study  
13 system. While sampling large number of genes has given us a tremendous opportunity to uncover  
14 previously unknown relationships and improve phylogenetic resolution, it also presents us with  
15 new challenges when the phylogenetic signal is confused by differences in the evolutionary  
16 histories of sampled genes. Given the addition of accurate marginal likelihood estimation methods  
17 into popular Bayesian software programs, it is natural to consider using the Bayes Factor (BF) to  
18 compare different partition models in which genes within any given partition subset share both  
19 tree topology and edge lengths. We explore using marginal likelihood to assess data subset  
20 combinability when data subsets have varying levels of phylogenetic discordance due to deep  
21 coalescence events among genes (simulated within a species tree), and compare the results with  
22 our recently-described phylogenetic informational dissonance index (D) estimated for each data  
23 set. BF effectively detects phylogenetic incongruence, and provides a way to assess the statistical  
24 significance of D values. We discuss methods for calibrating BFs, and use calibrated BFs to assess  
25 data combinability using an empirical data set comprising 56 plastid genes from green algae order  
26 Volvocales.

27 **Keywords:** Bayes Factor, concatenation, marginal likelihood, phylogenetic dissonance,  
28 phylogenetics, phylogenomics, Lindley's Paradox

29

## INTRODUCTION

30           Until recently, common practice for inferring multi-gene phylogenies involved  
31 concatenation of all available genes with an assumption that the evolutionary histories of all  
32 sampled genes were identical. However, phylogenetic trees for different genes (gene trees) can  
33 differ from each other, from the tree inferred from the concatenated data, and from the true  
34 species tree, due to evolutionary events/processes such as incomplete lineage sorting (ILS),  
35 horizontal transfer, and hybridization (Maddison, 1997; Edwards, 2009; Degnan and Rosenberg,  
36 2009; Mallet et al., 2016). Further, even if the sampled genes share the same evolutionary history,  
37 estimated trees can differ because of: (1) insufficient phylogenetic information in the sampled  
38 genes (stochastic or sampling error), or (2) model misspecification (systematic error) leading to,  
39 for example, long edge attraction in some gene trees and not in others (Swofford et al., 1996;  
40 Philippe et al., 2005, 2011).

41           With the recent surge of large-scale genomic DNA data from high-throughput sequencing  
42 methods, the issue of phylogenetic incongruence has become even more important in phylogeny  
43 reconstruction. Inferring species trees by addressing these challenges has become an area of active  
44 research in phylogenetics. Several species tree methods already available (reviewed in Liu et al.,  
45 2015) are effective in correcting incongruences due to deep coalescence (e.g. Song et al., 2012; Xi  
46 et al., 2014; Jarvis et al., 2014; Tang et al., 2015). These methods estimate a species tree either  
47 from multiple sequence alignments (e.g. \*BEAST, Heled and Drummond, 2010; BEST, Liu et al.,  
48 2008; SVDquartets, Chifman and Kubatko, 2014, 2015) or summary statistics calculated from

49 estimated gene trees (e.g. STEM, Kubatko et al., 2009; MP-EST, Liu et al., 2010; BUCKy, Ané  
50 et al., 2007; ASTRAL, Mirarab et al., 2014b). Methods such as \*BEAST and BEST  
51 simultaneously estimate gene trees and the species tree by using MCMC to integrate over trees  
52 and substitution model parameters; however, co-estimation of species and gene trees under a  
53 multispecies coalescent model is computationally intensive and cannot be applied to large scale  
54 genomic data. On the other hand, fast and efficient summary statistic methods (e.g. Mirarab  
55 et al., 2016) that completely rely on the estimated gene tree/trees (partial data) for the  
56 downstream species tree estimation may be prone to systematic bias as they do not incorporate  
57 uncertainty in the gene tree estimation process. Still lacking is a comprehensive approach that  
58 employs both a rigorous and more efficient algorithm to estimate species trees with high accuracy  
59 from hundreds of loci by addressing not just one (e.g. ILS) but all sources of phylogenetic  
60 incongruence (Posada, 2016). Until such methods are widely available, there is a need to at least  
61 identify phylogenetically congruent sets of loci among sampled genes. Phylogenies from congruent  
62 sets of genes may then be used to estimate a species phylogeny (cf. statistical binning, Mirarab  
63 et al., 2014a). Furthermore, identifying genes that are significantly incongruent may also be used  
64 to identify sequences resulting from processes other than the standard vertical inheritance model  
65 assumed in most phylogenetic analyses.

66 **Phylogenetic dissonance.**— Lewis et al. (2016) introduced Bayesian methods for  
67 measuring the phylogenetic information content of data and for measuring the degree of  
68 phylogenetic informational dissonance among data subsets. Phylogenetic dissonance is relevant to  
69 the problem of identifying congruent subsets of loci. When data are partitioned into subsets

70 (corresponding to, for example, genes or codon positions), such tools yield insight into which data  
71 subsets have the greatest potential for producing well supported estimates of phylogeny. Conflict  
72 between different subsets with respect to tree topology can lead to paradoxical results with  
73 respect to both information content and estimated phylogeny. For example, a tree topology  
74 minimally supported by all subsets (posterior probability less than 0.2) may be given maximal  
75 support (posterior probability 1.0) in a concatenated analysis if each subset is highly informative  
76 and effectively rules out the trees most supported by other subsets (Lewis et al., 2016). The  
77 information measure  $D$  (phylogenetic dissonance) was introduced by Lewis et al. (2016) to  
78 specifically identify such anomalies. Phylogenetic dissonance is defined as

$$\hat{D} = \hat{H}_{\text{merged}} - \hat{H}_{\text{average}} \quad (1)$$

$$\hat{H}_{\text{average}} = \frac{1}{K} \sum_{k=1}^K \hat{H}_k, \quad (2)$$

79 where  $\hat{H}_k$  is the entropy of the marginal tree topology posterior distribution for data subset  $k$  (of  
80  $K$  subsets), and  $\hat{H}_{\text{merged}}$  is the entropy of a posterior distribution estimated from a merged tree  
81 sample. Posterior tree samples from separate analyses of each data subset are combined to form  
82 the merged tree sample. (Note that this merged tree sample differs from a tree sample obtained  
83 from a concatenated analysis.) If different data subsets strongly support mutually exclusive tree  
84 topologies, then the average entropy of marginal tree topology posterior distributions ( $\hat{H}_{\text{average}}$ )  
85 will be small while the merged entropy ( $\hat{H}_{\text{merged}}$ ) will be relatively large due to the fact that  
86 topology frequencies are more evenly distributed in the merged sample compared to samples from  
87 individual subsets, which are each dominated by one tree topology. Lewis et al. (2016) defined  
88 and estimated phylogenetic dissonance using this entropy-based measure, but how to evaluate the

89 statistical significance of a given level of phylogenetic dissonance remains an open question.

90           **Tests for Phylogenetic Dissonance.**— The only direct tests of phylogenetic  
91 congruence proposed to date are likelihood ratio tests (LRTs). Huelsenbeck and Bull (1996)  
92 proposed a parametric bootstrapping approach in which the null hypothesis constrained all data  
93 subsets to have the same tree topology, while the alternative (unconstrained) hypothesis allowed  
94 each subset to have a potentially different tree topology. The distribution of the test statistic was  
95 generated by simulating data sets under the null hypothesis using maximum likelihood estimates  
96 of all model parameters and computing the test statistic under each simulated data set.

97           Non-parametric bootstrapping, in conjunction with LRTs, was used by Leigh et al. (2008)  
98 to test the same null hypothesis. Leigh et al. (2008) also proposed clustering of data subsets based  
99 on pairwise LRT results to generate compatible sets. Separate likelihood ratio tests were also  
100 proposed by Leigh et al. (2008) to test for heterotachy: in this case the null hypothesis constrains  
101 edge lengths to be proportionally identical across subsets, while the alternative hypothesis allows  
102 each subset to potentially have different edge lengths. The software CONCATERPILLAR (Leigh  
103 et al., 2008) may be used to carry out these non-parametric bootstrapping LRTs.

104           These likelihood ratio tests are well justified and are the best available means to assess  
105 congruence when there are no priors involved in the tree estimation process. However, when the  
106 phylogeny estimation involves Bayesian methods, then evaluation of congruence should properly  
107 account for the effects of the assumed prior distributions. We propose a Bayesian approach to  
108 testing phylogenetic congruence (or, equivalently, dissonance) by comparing the marginal

109 likelihoods of competing models. When only two models are compared, the ratio of marginal  
110 likelihoods is termed the Bayes Factor (BF). Our approach is comparable to that of Leigh et al.  
111 (2008), but instead of comparing maximized log-likelihoods of competing models using LRTs, we  
112 use marginal likelihoods and their ratio (BF) for model comparison. Our approach is made  
113 possible by the recent improvements in marginal likelihood estimation (stepping-stone, SS: Xie  
114 et al., 2010, Fan et al., 2011; path-sampling, PS: Lartillot and Philippe, 2006; partition weighted  
115 kernel estimator, PWK: Wang et al., 2017) for phylogenetic model selection. The SS and PS  
116 estimators substantially outperformed other approaches (e.g. harmonic mean estimator, HME,  
117 and a posterior simulation-based analog of Akaike's information criterion through Markov chain  
118 Monte Carlo, AICM) for comparing models of demographic change and relaxed molecular clocks  
119 (Baele et al., 2012). Recently, Brown and Thomson (2016) also used BF to analyze the sensitivity  
120 in clade resolution to the data types used to infer the topology. The primary aim of our study is  
121 to evaluate the effectiveness of BF for assessing significance of the phylogenetic dissonance  
122 measure  $D$  (equation 1). We explore the behavior of BF using simulations designed to create a  
123 spectrum of 10-gene data sets ranging from low to high information content and from complete  
124 topological concordance to extreme discordance (due to deep coalescence and subsequent  
125 incomplete lineage sorting). We also provide an empirical example involving concordance of  
126 nuclear and plastid genes in the green algal order Volvocales which demonstrates that likelihood  
127 ratio tests carried out using CONCATERPILLAR can differ from conclusions based on marginal  
128 likelihoods when analyses are performed in a Bayesian context.

129

## MATERIALS AND METHODS

130 **Bayes Factors.**— In Bayes' Rule,

$$p(\tau, \phi_M | \mathbf{y}, M) = \frac{p(\mathbf{y} | \tau, \phi_M, M) p(\phi_M | \tau, M) p(\tau | M)}{\sum_{\tau} p(\tau | M) \int p(\mathbf{y} | \tau, \phi_M, M) p(\phi_M | \tau, M) d\phi_M},$$

131 the denominator represents the marginal likelihood  $p(\mathbf{y} | M)$ : the total probability of data  
132  $\mathbf{y}$  given model  $M$ , averaged over tree topology  $\tau$  and a multivariate parameter vector  $\phi_M$   
133 comprising model parameters. The parameters composing  $\phi_M$  may be tree-specific (e.g. edge  
134 lengths) or substitution-model-specific (e.g. transition/transversion rate ratio). Data  $\mathbf{y}$  is a vector  
135 comprising observed patterns of states for all taxa for individual characters (sites in the case of  
136 sequence data). Considering two models,  $(M_1, M_2)$ , and their marginal likelihoods,  $p(\mathbf{y} | M_1)$  and  
137  $p(\mathbf{y} | M_2)$ , respectively, the BF  $B_{12}$  is the ratio  $p(\mathbf{y} | M_1) / p(\mathbf{y} | M_2)$ . The BF on the log-scale is  
138 calculated as:

$$\log B_{12} = \log p(\mathbf{y} | M_1) - \log p(\mathbf{y} | M_2),$$

139 where  $\log B_{12} > 0$  signifies that model  $M_1$  is preferred over  $M_2$ . By preferred, we mean that  
140 model  $M_1$  fits the data better on average than model  $M_2$  over the parameter- and tree-space  
141 defined by the prior. Applying this approach to the problem of phylogenetic congruence, consider  
142 data from a set of  $K$  loci  $\mathbf{y}$  ( $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_K$ ), and two models, CONCATENATED and SEPARATE.  
143 The CONCATENATED model represents the marginal likelihood of the concatenated set ( $\mathbf{y}_C$ ) in  
144 which all loci are forced to have the same topology and model parameters ( $\phi_M$ ),

$$p(\mathbf{y} | M = \text{CONCATENATED}) = \sum_{\tau} p(\tau) \int p(\mathbf{y}_C | \tau, \phi_M) p(\phi_M) d\phi_M, \quad (3)$$



145 whereas the SEPARATE model represents the marginal likelihood for a model in which individual  
 146 loci are allowed to have their own topologies and model parameters  $(\phi_{M_1}, \phi_{M_2}, \dots, \phi_{M_K})$ ,

$$p(\mathbf{y}|M = \text{SEPARATE}) = \prod_{k=1}^K \left( \sum_{\tau_k} p(\tau_k) \int p(\mathbf{y}_k|\tau_k, \phi_{M_k}) p(\phi_{M_k}) d\phi_{M_k} \right). \quad (4)$$

147 The BF for CONCATENATED against SEPARATE is defined

$$B_{CS} = \frac{p(\mathbf{y}|M = \text{CONCATENATED})}{p(\mathbf{y}|M = \text{SEPARATE})}.$$

148 When the tree topology prior is discrete uniform,

$$B_{CS} = \frac{N_T^{K-1} \sum_{\tau} \int p(\mathbf{y}_C|\tau, \phi_M) p(\phi_M) d\phi_M}{\prod_{k=1}^K \left( \sum_{\tau_k} \int p(\mathbf{y}_k|\tau_k, \phi_{M_k}) p(\phi_{M_k}) d\phi_{M_k} \right)},$$

149 where  $N_T$  equals the number of distinct labeled tree. Here,  $B_{CS} > 1$  (or equivalently  $\log B_{CS} > 0$ )  
 150 indicates that the CONCATENATED model (numerator) is preferred over the SEPARATE  
 151 model (denominator), whereas  $B_{CS} < 1$  ( $\log B_{CS} < 0$ ) indicates the reverse (i.e. SEPARATE  
 152 model is the preferred model).

A third, intermediate model HETERO links topology across subsets but allows edge lengths to vary between single-gene data sets:

$$p(\mathbf{y}|M = \text{HETERO}) = \sum_{\tau} p(\tau) \prod_{k=1}^K \left( \int p(\mathbf{y}_k|\tau, \phi_{M_k}) p(\phi_{M_k}) d\phi_{M_k} \right).$$

153 While BF may be defined between any pair of models, and while we continue to describe our  
 154 approach as using Bayes Factors, in practice we will only implicitly compute BF, instead  
 155 estimating the log marginal likelihood of each of the three models and declare the winning model  
 156 as the one having the largest of the three log marginal likelihood values.

157           **Data Simulation.**— Gene trees were simulated within species trees using parameter  
158 combinations that yielded differing levels of phylogenetic incongruence. Using a Python script  
159 (source code provided in Supplementary Materials), one thousand 6-taxon species trees were  
160 generated under a pure-birth (Yule) process in which the tree height  $T$  (expected number of  
161 substitutions along a single path from root to tip) was drawn from a Lognormal(0.05, 0.22)  
162 distribution (mean 1.08, 95% of samples between 0.68 and 1.62). Ten gene trees were simulated  
163 within each species tree using coalescent parameter  $\theta = 4N_e\mu$ , where  $N_e$  is the effective (diploid)  
164 population size and  $\mu$  is the mutation rate per generation. For each species tree, the ratio  $\theta/T$   
165 was drawn from a Lognormal(0.60, 0.77) distribution (which has mean 2.45 with 95% of samples  
166 between 0.40 and 8.24) and  $\theta$  was determined by multiplying this ratio by the value of  $T$  used for  
167 a specific species tree. Increasing  $\theta$  relative to  $T$  results in a higher number of deep coalescences,  
168 causing increased discordance among the gene trees.

169           The gene trees thus generated were subsequently used to simulate DNA sequence  
170 alignments of length 2000 sites using seq-gen (Rambaut and Grass, 1997) under the HKY+G  
171 model. Individual single-gene datasets and the concatenated dataset were used to compute  
172 marginal likelihoods using the Stepping-stone method (Xie et al., 2010) implemented in MrBayes  
173 (Ronquist et al., 2012). For the concatenated dataset, two marginal likelihoods were estimated by  
174 enforcing: (1) the same topology and edge lengths for all sites (CONCATENATED model), and  
175 (2) the same topology but allowing edge lengths to vary among single-gene data subsets to  
176 account for non-topological gene tree variation (HETERO model). Analyses of single-gene data  
177 sets alone yielded marginal likelihoods that, when multiplied together, yield the marginal

178 likelihood under the SEPARATE model.

179           In order to assess the robustness of BF for detecting topological and edge length  
180 congruence, the BF results were evaluated with respect to the phylogenetic information content  
181 ( $I$ ) and phylogenetic dissonance ( $D$ ) values computed using Galax v1.0.0 (Lewis et al., 2016).  
182 Estimation of  $I$  and  $D$  uses conditional clade probabilities (Larget, 2013) to estimate Shannon  
183 entropy (Shannon, 1948), from which  $\hat{I}$  is calculated simply as a difference between the entropies  
184 of the marginal prior and marginal posterior distributions of tree topology (Lindley, 1956). The  
185 phylogenetic dissonance is defined as in equation (1), and thus  $\hat{D}$  is computed as the entropy of  
186 the merged tree sample minus the average entropy of tree samples from individual genes. We also  
187 tested the strength of different variables including  $\hat{D}$  (and their combinations) in discriminating  
188 SINGLE vs. CONCATENATED model by conducting a linear discriminant analysis (LDA). The  
189 LDA was carried out in R using the 'lda' function available in the library MASS (Venables and  
190 Ripley, 2002) for all the predictor variables (number of conflicting nodes, number of variable sites,  
191 number of parsimony informative sites,  $\theta/T$ , species tree height/shortest gene tree height, species  
192 tree height/longest gene tree height, average information content,  $D$ , and number of deep  
193 coalescences).

194           Phylogenetic dissonance is expected to be zero for comparisons of independent MCMC  
195 samples from the same posterior distribution, and thus provides a sensitive measure of MCMC  
196 convergence with respect to tree topology (Lewis et al., 2016). We replicated each single-gene and  
197 concatenated MCMC analysis and computed  $\hat{D}$  for these paired samples as a way of ensuring that  
198 post burn-in MCMC sample size was sufficient for convergence.

199           **Lindley’s Paradox.**— The tendency of Bayes Factors to prefer a sharp null hypothesis  
200 (e.g. a point mass prior) over an a priori diffuse (e.g. noninformative) alternative hypothesis  
201 when a classical frequentist hypothesis test would reject the null hypothesis is known as Lindley’s  
202 Paradox (Jeffreys, 1939; Lindley, 1957). The BF is identical to the posterior model odds given  
203 equal model prior probabilities. Giving both the sharp null hypothesis and the diffuse alternative  
204 hypothesis equal prior weight provides a distinct advantage for the null hypothesis as long as the  
205 null hypothesis represents a better explanation of the data compared to most parameter values  
206 supported by the alternative hypothesis. The amount of this advantage grows with the a priori  
207 diffuseness of the alternative hypothesis.

208           Consider the BF for CONCATENATED against SEPARATE models. Equation (3) shows  
209 that the marginal likelihood of the CONCATENATED model contains a term  $p(\tau)$  that equals  
210 the prior probability of the tree topology shared among all data subsets. Assuming a discrete  
211 uniform prior distribution over tree topologies,  $p(\tau)$  is a constant equal to  $1/N_T$ . Equation (4)  
212 shows that the corresponding term in the marginal likelihood for the SEPARATE model is  
213  $(1/N_T)^K$ , reflecting the fact that each of the  $K$  genes potentially has a different tree topology. As  
214 either  $N_T$  or  $K$  increases, the CONCATENATED model becomes increasingly sharp compared to  
215 the SEPARATE model with respect to prior distributions and thus Lindley’s paradox must be  
216 taken into consideration given a sufficiently large number of taxa and/or data subsets. In other  
217 words, for large trees or large number of genes, or both, assuming a common tree for all genes  
218 may provide a better explanation, even if incorrect in some details, than allowing each gene to  
219 have its own tree topology (and independent set of edge lengths). Here, model *fit* is viewed from

220 the Bayesian perspective and is thus more appropriately described as *average fit*. It is the fact  
221 that model fit is averaged over a very large number of incorrect trees, each considered equal by  
222 the prior, that drags down the marginal likelihood of the SEPARATE model.

223       Using BF for testing data combinability must keep the possibility of Lindley's Paradox in  
224 mind. Fortunately it is not difficult to determine if Lindley's Paradox applies: if the likelihood  
225 ratio test approach chooses the SEPARATE model but BF chooses CONCATENATE, this  
226 provides a strong hint that it is the vagueness of the prior in the SEPARATE model that is  
227 tipping the balance. While this is less a paradox than a difference in Bayesian vs. Frequentist  
228 perspective, a researcher may nevertheless wish to lessen the impact of the tree topology prior on  
229 the model choice decision.

230       While the prior distributions for edge lengths and substitution model parameters are  
231 potentially relevant to Lindley's paradox, these parameters are not directly involved in the test  
232 and are integrated out of both numerator and denominator in the BF calculation. Bergsten et al.  
233 (2013) identified similar issues related to diffuse tree topology priors in BF used for testing  
234 monophyly. In that case, constraints placed on tree topologies to enforce monophyly affect the  
235 size of tree space, which creates an imbalance in tree topology priors analogous to that  
236 encountered when testing for data combinability.

237       **BF Calibration.**— It is standard practice to use the value  $BF = 1$  as the critical value  
238 determining whether the null model (e.g. CONCATENATED) or the alternative model (e.g.  
239 SEPARATE) wins. This makes sense when the prior predictive error probabilities of BF under

240 both models are equal; however, in cases where models differ substantially in their effective  
241 dimensions, the distributions of the BF for the two models being compared may not be  
242 symmetrical. For example, it is possible that the probability of choosing the CONCATENATED  
243 model when the SEPARATE model is true may not equal the probability of choosing the  
244 SEPARATE model when the CONCATENATED model is true:

$$p(B_{CS1}|\text{SEPARATE}) \neq p(B_{CS1}|\text{CONCATENATED}).$$

245 Under such circumstances, a different threshold value (other than 1) can be selected such that the  
246 probability of choosing the incorrect model under both hypotheses is equal. García-Donato and  
247 Chen (2005) suggested a method for calibrating the BF that makes the prior predictive error  
248 probabilities symmetrical. To apply the method of García-Donato and Chen (2005), we simulated  
249 1000 replicate 6-taxon, 10-gene data sets (2000 sites/gene) from the joint prior distribution of  
250 each model (CONCATENATED and SEPARATE). For the CONCATENATED model, data for  
251 all 10 genes were simulated from a single topology sampled from the discrete uniform topology  
252 prior. For the SEPARATE model, data for each of the 10 genes was simulated from topologies  
253 separately sampled from the discrete uniform topology prior. All other model parameters were  
254 simulated from their respective prior probability distributions.

255 For each simulated data set,  $B_{CS}$  was computed, yielding a sample of 1000 values from  
256 the prior predictive BF distributions for both the CONCATENATED and SEPARATE models.  
257 The 2000 sampled BF values were combined into a single vector and sorted, and the critical value  
258  $c$  was chosen as the midpoint between the 1000th and 1001th values in the sorted vector. This

259 procedure identifies a BF cutoff value  $c$  that satisfies

$$p(B_{CSc}|\text{SEPARATE}) = p(B_{CSc}|\text{CONCATENATED}).$$

260 The simulations needed for BF calibration were carried out using PAUP\* 4a158 (Swofford, 2003).

261       **Example from the Green Algal Order Volvocales.**— We tested phylogenetic  
262 congruence among 56 protein-coding plastid genes used in Fučíková et al. (2016), focusing on one  
263 of the most topologically consistent parts of the tree, the green algal order Volvocales. The  
264 Volvocales dataset consisted of a subset of the Sphaeropleales, Vovocales, and OCC  
265 (Oedogoniales- Chaetopeltidales- Chaetophorales) clades studied in Fučíková et al. (2016). We  
266 included four of the five Volvocales members from the study: *Chlamydomonas reinhardtii*,  
267 *Gonium pectoral*, *Pleodorina starrii*, and *Volvox carteri*. The length of post-trimmed plastid genes  
268 ranged from 93 sites (*psbT*) to 2259 sites (*psaA*). We conducted BF tests for all possible pairs  
269 from the 56 genes (by estimating marginal likelihoods under the CONCATENATED and  
270 SEPARATE models) used in the study with the aim to detect possible outlier genes that may be  
271 present among the sampled genes for the concatenated phylogeny. The critical value  $c$  for this  
272 analysis was computed using the same approach as simulated data. The prior predictive  
273 distributions of BF under CONCATENATED and SEPARATE models were obtained from 1000  
274 replicates (4-taxon, 2 genes/replicate, and 2000 sites/gene) simulated under each model using  
275 PAUP 4a158 (Swofford, 2003). For the CONCATENATED model, DNA sequence data for both  
276 genes were simulated from a single topology (randomly drawn from the discrete uniform topology  
277 prior) with edge lengths and other model parameters drawn from the GTR+G model prior  
278 distribution, whereas for the SEPARATE model, sequence data for each of the 2 genes were

279 simulated from individually drawn discrete-uniform-distributed topologies with all other model  
280 parameters drawn from the GTR+G model prior distribution. In order to compare our results  
281 with the likelihood-based approach, we also tested congruence among these 56 genes using  
282 CONCATERPILLAR (Leigh et al., 2008) using its topological congruence test (-t) option.

Parameters of the models and the priors used in the study of simulated (model: HKY+G)  
and Volvocales (model: GTR+G) data were:

Tree topology  $\tau \sim \text{Discrete Uniform}(1, T)$

Tree length  $L \sim \text{Exponential}(0.1)$

Edge length proportions  $\mathbf{e} \sim \text{Dirichlet}(1, \dots, 1)$

Nucleotide frequencies  $\boldsymbol{\pi} \sim \text{Dirichlet}(1, 1, 1, 1)$

transition/transversion rate ratio  $\kappa \sim \text{Beta}(1, 1)$

Exchangeabilities  $\mathbf{r} \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)$

Discrete Gamma shape  $\alpha \sim \text{Exponential}(1)$ ,

283 where  $T$  equals the total number of distinct, labeled, binary, unrooted tree topologies.

## 284 RESULTS

285 **Bayes Factor Calibration For Simulation Study.**— BF calibration for the 6-taxon,  
286 10-gene simulation study resulted in a critical value  $c = -3.2$  (log scale), which equals the



287 midpoint in the interval extending from the 1000<sup>th</sup> element (-3.55) to the 1001<sup>st</sup> element (-2.85) of  
288 the combined, sorted vector of prior predictive log  $B_{CS}$  values from CONCATENATED and  
289 SEPARATE models. Using the standard log-BF cutoff (0.0) would thus result in the SEPARATE  
290 model winning more often when CONCATENATED is the true model than the  
291 CONCATENATED model wins when the SEPARATE model is true.

292 **Phylogenetic Dissonance Correlated with Number of Deep Coalescences.**— As  
293 expected, estimated phylogenetic dissonance ( $\hat{D}$ ) was correlated with number of deep coalescences  
294 in 1000 simulated gene sets (10 genes/set) representing various degrees of topological and edge  
295 length congruence (Fig. 2). The number of deep coalescences varied from the minimum possible  
296 number (0) to the maximum possible number (50). (The maximum number of deep coalescences  
297 is 5 per gene because there are 5 internal nodes in a rooted tree of 6 taxa.)

298 In our simulations, under both criteria ( $c = 0$ ,  $c = -3.2$ ), the SEPARATE model won in a  
299 majority of replicates when  $\hat{D} > 1.2$  or when the number of deep coalescences exceeded 1.8 per  
300 gene. Under the new critical value ( $c = -3.2$ ), 1 simulation replicate switched its support to the  
301 CONCATENATED model from the earlier SEPARATE or HETEROTACHY model. When  
302 SEPARATE failed to have the largest log marginal likelihood, CONCATENATED usually won,  
303 with HETERO only achieving the largest log marginal likelihood if  $\hat{D} < 1.2$  and the number of  
304 deep coalescences was less than 3.2 per gene.

305 In cases of multiple deep coalescences ( $>3.2$ /gene) or high dissonance  $\hat{D} > 1.3$ , the  
306 CONCATENATED model won only when average information content was low, while HETERO

307 never won under these circumstances. In a few cases, SEPARATE was the winning model even  
308 when the number of deep coalescences was less than 1 per gene on average. Conversely,  
309 CONCATENATED was occasionally the winner despite high levels of deep coalescence ( $> 3.5$  per  
310 gene).

311

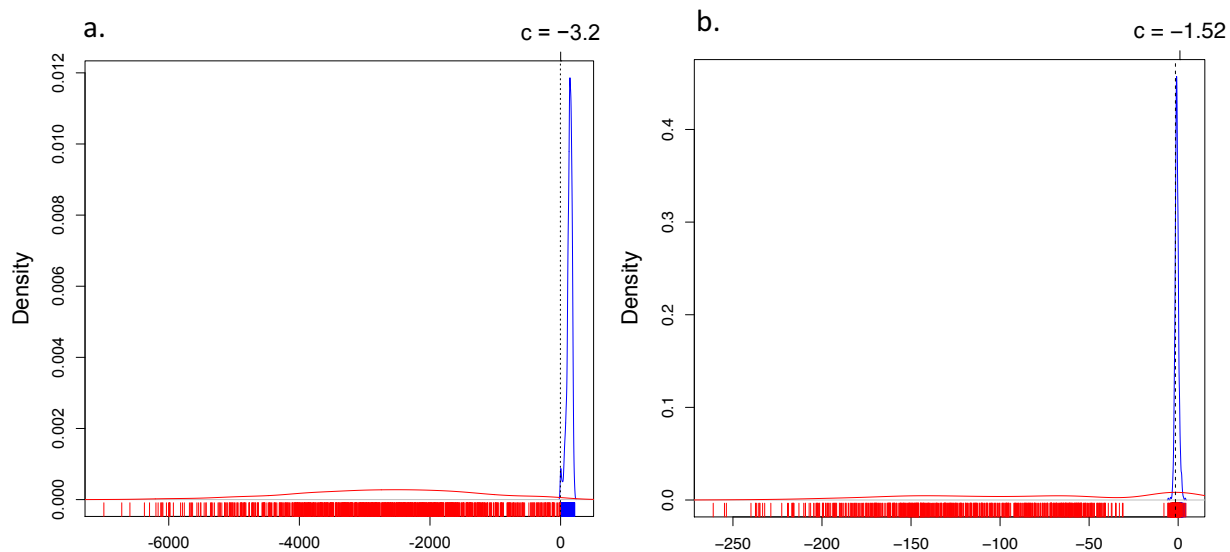


Figure 1: Density of  $B_{CS}$  under CONCATENATED (blue line) and SEPARATE (red line) models for the (a) 6 taxa, 10-gene data set and (b) 4 taxa 2-genes data set. The critical values ( $c = -3.2$ ,  $c = -1.52$ ) are indicated by dashed lines, estimated using 1000 prior predictive replicates from each model (rug).

312

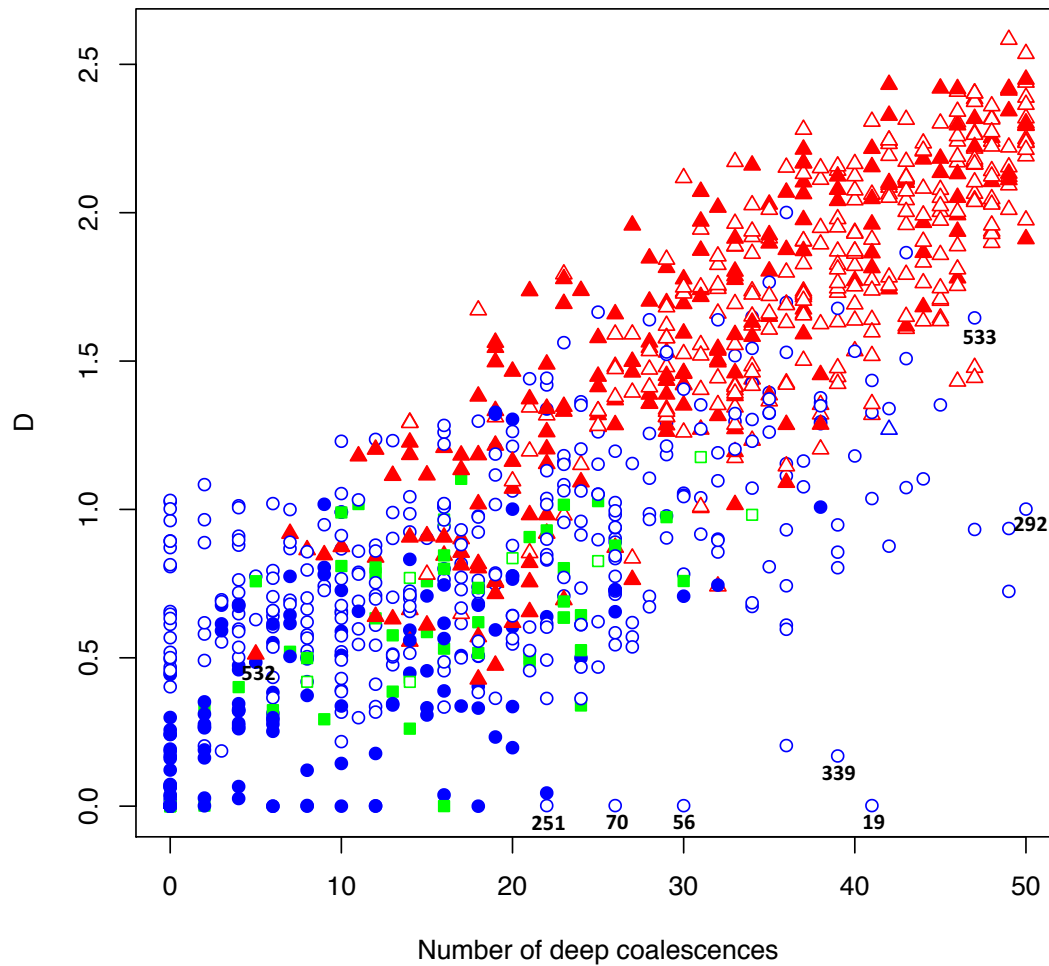


Figure 2: Plot of 1000 replicates simulated under conditions that yielded varying levels of deep coalescence (x-axis) and phylogenetic dissonance  $\hat{D}$  values (y-axis). Blue circles indicate Bayes Factor support for the CONCATENATED model over both HETERO and SEPARATE with log-scale critical value  $c = 0$ . Blue triangles indicate Bayes Factor support for the CONCATENATED model over both HETERO and SEPARATE with  $-3.2 < c < 0$ . Green square indicates support for the HETERO model over both CONCATENATED and SEPARATE. Red triangle indicates support for SEPARATE model over both CONCATENATED and HETERO with  $c = -3.2$ . Filled symbols represent  $\geq 90\%$  average information content across genes, with closed symbols indicating  $< 90\%$ . Numbers indicate particular replicates mentioned in the text.

313

314           **Volvocales Example.**— The results of the pair-wise tests of congruence are illustrated  
315 in Fig. 3a. The critical value  $c$  computed for the four-taxon case based on the prior predictive  
316 distributions of BF under CONCATENATED and SEPARATE models was -1.52 (Fig. 1b).  
317 Under both criteria ( $c = 0$ ,  $c = -1.52$ ), marginal likelihoods indicated congruence for all gene pairs  
318 with the exception of *petD* and *rpl36*, each of which was incongruent with every other gene (but  
319 were congruent with each other). Both *petD* and *rpl36* favor *Gonium + Pleodorina* (Fig. 3b) while  
320 all other genes favor *Volvox + Pleodorina* (Fig. 3c). The CONCATERPILLAR analysis, however,  
321 indicated that all 56 genes were topologically congruent. The two genes found to be incongruent  
322 using BF analysis (*petD* and *rpl36*) were not contiguous in the chloroplast genomes of four taxa,  
323 suggesting that they were not the result of a single horizontal transfer event. In the case of *petD*,  
324 there is a single variable amino-acid site (amino-acid position 106) that determines the *Gonium +*  
325 *Pleodorina* relationship. Excluding site 106 removes support for this relationship. Despite the  
326 incongruence of *rpl36* to the other genes, this particular gene is short (total nucleotide length  
327 =114) and it contains relatively less information relevant to estimating topology. We also used  
328 PhyloBayes (Lartillot et al., 2009) to estimate phylogeny for the *petD* data (including all the  
329 sequences from Chlorophyceae available on Dryad: <http://dx.doi.org/10.5061/dryad.q8n0v>)  
330 under the CAT model (Lartillot and Philippe, 2004). The CAT model accommodates sites with  
331 distinct state frequency profiles, unlike standard models that assume state frequencies are  
332 homogeneous across sites. The CAT model can potentially avoid long-branch attraction due to  
333 the model assuming a wider range of available amino acids at particular sites than are actually

334 available. However, even under the CAT model, *Pleodorina* resolved sister to *Gonium*.

335

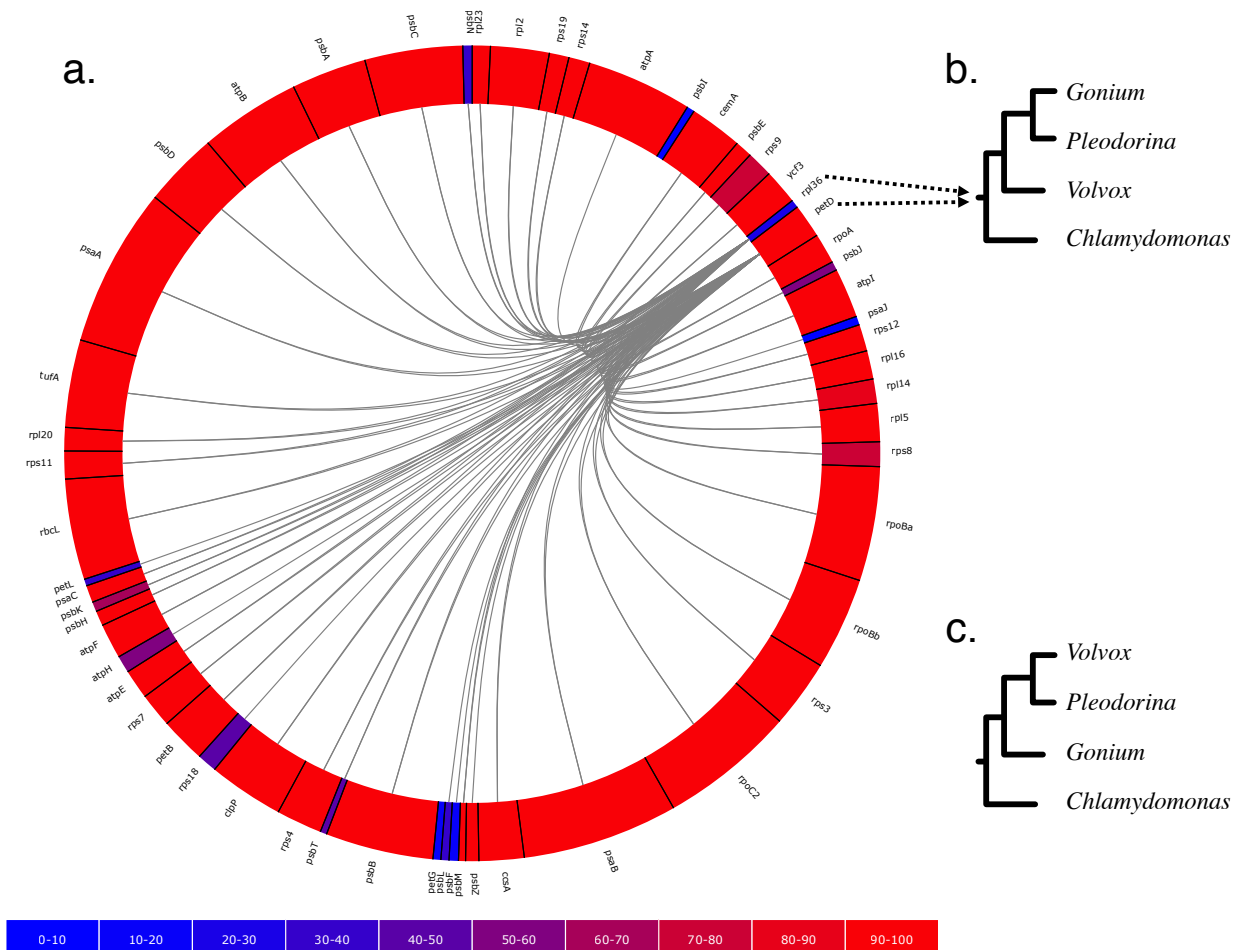


Figure 3: Pair-wise BF test for phylogenetic congruence for all possible pairs among 56 protein coding plastid genes (Fučíková et al., 2016) where the colors represent the information content present in the gene and the lines between the genes indicate phylogenetic incongruence (i.e. support for SEPARATE over CONCATENATED) suggested by the BF test (3a). In the 56 gene sets, *petD* and *rpl36* show support for *Gonium* + *Pleodorina* relationship (3b), whereas the other 54 genes support *Volvox* + *Pleodorina* relationship (3c).

336

337

## DISCUSSION

338           The presence of deep coalescence does not guarantee that different genes will have  
339 different tree topologies, but the fact that lineages are joined randomly when there is deep  
340 coalescence means that greater incongruence is the expected result of increasing the frequency of  
341 deep coalescence. In general, more deep coalescences yielded higher  $\hat{D}$  and a greater chance of the  
342 SEPARATE model winning. In fact,  $D$  was the most important variable in discriminating  
343 SINGLE vs. CONCATENATED model in the discriminant function analysis involving a number  
344 of other variables tested (number of conflicting nodes, number of variable sites, number of  
345 parsimony informative sites,  $\theta/T$ , species tree height/shortest gene tree height, species tree  
346 height/longest gene tree height, average information content,  $D$ , and number of deep  
347 coalescences). The  $D$ , number of deep coalescences, and  $\theta/T$  could separate SINGLE vs.  
348 CONCATENATED models with 91% accuracy where the  $D$  alone could separate the two with  
349 82% accuracy. Because a single simulation study can only suggest appropriate cutoff values for  $\hat{D}$   
350 for the limited range of parameter combinations explored, we argue that a BF approach provides  
351 a sensible general approach for determining when values of  $\hat{D}$  are too high to be compatible with  
352 phylogenetic congruence.

353           It is interesting and informative to examine some outliers in the simulation results  
354 presented in Figure 2. For example, consider replicate 532, for which the SEPARATE model won  
355 despite a high average information content (94% of maximum information), relatively low  $D$ , and  
356 a single topological conflict among 10 genes. Removing the gene that conflicted (gene2) from the

357 concatenated set, followed by re-estimation of marginal likelihoods, resulted in a win for the  
358 CONCATENATED model, suggesting that a single incongruent subset out of 10 total can be  
359 enough to place the SEPARATE model on top.

360         Low phylogenetic signal can result in a preference for the CONCATENATED model  
361 despite a high number of deep coalescences (e.g. Fig. 2, replicates 19, 56, 70, 97, 251, 292, 339,  
362 533). In some extreme cases, when phylogenetic information content is very low (approaching  
363 zero information),  $\hat{D}$  can also be low (Fig. 2, replicates 19, 56, 70, 251). In such cases, posterior  
364 samples of individual gene subsets visit every possible tree topology (of the 105 possible unrooted  
365 binary tree topologies for 6 taxa) in roughly equal proportions. Phylogenetic dissonance is zero if  
366 all subset posterior distributions are equal, and this is true whether these posterior distributions  
367 are concentrated or flat, so low  $\hat{D}$  in the face of low information content for all gene subsets is not  
368 surprising. It is also unsurprising that the marginal likelihood would favor the  
369 CONCATENATED model in such cases because one tree topology is about as good as any other  
370 tree topology in explaining the data, and the marginal likelihood implicitly punishes models for  
371 including parameters that do not provide access to regions of parameter space providing  
372 appreciably higher likelihood.

373         **The Case of Mistaken Heterotachy.**— An interesting phenomenon was observed as  
374 a result of using phylogenetic dissonance to assess MCMC convergence with respect to tree  
375 topology. Most replicate MCMC analyses exhibited  $\hat{D} < 0.1$ , indicating that the posterior  
376 samples from replicate analyses were essentially identical (as they should be if both Markov

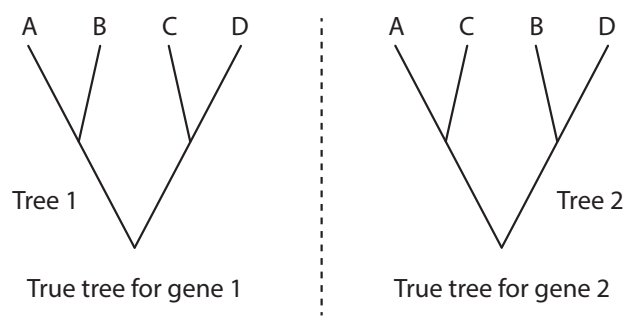
377 chains mixed well and were sampled only after converging to the stationary distribution);  
378 however, many analyses exploring the HETERO model produced unexpectedly high replicate  
379 phylogenetic dissonance values. The reason for this turns out to be the completely  
380 understandable result of a model making the best of a bad situation, and offers a warning for  
381 those who might be tempted to use a HETERO model win as an evidence for heterotachy.

382         Consider a case of two data subsets (genes) in which the true tree topology differs (Fig. 4).  
383 The HETERO model assumes that the same tree topology applies to both genes (which is *not*  
384 true in this case), but allows each gene to have its own set of edge lengths. The HETERO model  
385 can choose to focus on the true tree topology for gene 1 and attempt, using edge lengths, to  
386 explain the data for gene 2 as best it can. Alternatively, it can focus on the true tree topology for  
387 gene 2 and attempt, using edge lengths, to explain the data for gene 1 to the extent possible. How  
388 does a model fit data when assuming an incorrect tree topology? The answer is that it increases  
389 the lengths of edges for some taxa that are sister taxa in truth but not in the assumed tree,  
390 leaving other closely related taxa connected by relatively short paths. Thus, the fact that some  
391 taxa are more similar than the tree topology suggests can be explained by the model using  
392 evolutionary convergence (the long edged taxa), while similarities between other taxa that seem  
393 far apart on the assumed tree topology are explained by a lowered rate of substitution. In  
394 replicate analyses, it is possible for one run to choose the tree topology for gene 1 and the other  
395 replicate to choose the tree topology for gene 2, yielding posterior distributions that are  
396 concentrated on conflicting tree topologies, which in turn produces high estimated phylogenetic  
397 dissonance. The lesson to be learned from this study is that a win by the HETERO model may



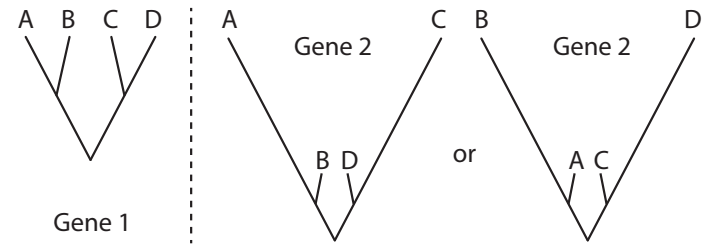
398 not mean the presence of heterotachy in data, but may simply reflect a model doing its best to  
399 explain data generated on a different tree topology. This crafty use of spurious edge lengths by  
400 models to explain away topological discordance among genes was explored in detail by Mendes  
401 and Hahn (2016). In their study of simulated and empirical data, Mendes and Hahn (2016) found  
402 that the topological discordance between gene trees due to ILS can cause multiple apparent  
403 substitutions on the focal tree (e.g. species tree) on one or more of its branches that uniquely  
404 define a split on the discordant gene tree that is absent in the species/focal tree. It is interesting  
405 that measuring phylogenetic dissonance among replicate analyses under the CONCATENATED  
406 model alone can potentially be used to detect incongruence in gene tree topologies.

407         The presence of true heterotachy is suggested by low phylogenetic dissonance combined  
408 with HETERO model being the winning model. None of our simulations imparted true  
409 heterotachy; however, some results (e.g. replicate 942) did combine  $\hat{D} = 0$  with a winning  
410 HETERO model. The explanation is that the HETERO model is actually detecting *heterochrony*  
411 (a new term) rather than *heterotachy*. Heterochrony may be defined as differences in the same  
412 edge length (measured in expected number of substitutions per site) across genes due to the fact  
413 that coalescence depth varies among genes (even if the topology is identical). The HETERO  
414 model is, in this case, detecting differences in coalescence times instead of differences in rate of  
415 substitution.

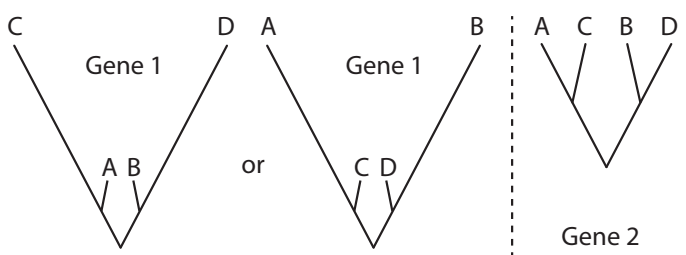


Suppose model forces both genes to have same topology, but each gene is allowed to have its own set of edge lengths.

Option 1: choose tree 1 and do your best to explain gene 2



Option 2: choose tree 2 and do your best to explain gene 1



Dissonance for 2 independent MCMC analyses will be large if each analysis, by chance, chooses a different focal tree topology

416

Figure 4: Explanation of paradoxical high dissonance for samples from independent replicate MCMC analyses exploring the same posterior distribution under the HETERO model.

417 **Empirical Volvocales Example.**— Our empirical example involved a reanalysis of a  
 418 subset of four taxa from a more inclusive study of green algal phylogeny. In that former study,

419 Fučíková et al. (2016) found strong support for a single tree topology relating these four taxa  
420 using a concatenated dataset, but reported very low internode certainty (IC: Salichos et al., 2014)  
421 values for all but one edge in the estimated tree. This suggests some conflict exists among genes,  
422 and thus it is not surprising that our BF analyses identified two genes (*rpl36* and *petD*) that  
423 preferred a different tree topology than the majority (54/56) of genes. What is perhaps surprising  
424 is that likelihood ratio tests conducted using CONCATERPILLAR found no conflict, concluding  
425 that all 56 genes should be concatenated. The fact that our BF approach and  
426 CONCATERPILLAR's LRT approach provide conflicting advice highlights a major difference  
427 between the Bayesian and frequentist statistical approaches to phylogenetics. For the *petD* gene,  
428 we found that a single amino acid site (site 106) determines the preference of this gene for  
429 *Gonium* + *Pleodorina*. Bayesian analyses do not take into consideration (either implicitly or  
430 explicitly) any data other than what was observed, and thus will take the evidence from site 106  
431 at face value. Assuming a site appears (to the model) to be a reliable reporter (i.e. substitution is  
432 rare and the site is not contradicted by any other site), then even one site may have a strong  
433 impact on a Bayesian phylogenetic analysis. Frequentist approaches involving bootstrapping,  
434 however, take additional sources of uncertainty into consideration. Bootstrapping evaluates many  
435 data sets, each different than the observed data set, and thus takes uncertainty in the observed  
436 data into account. This is one explanation for why bootstrap support values for clades tend to be  
437 smaller than posterior probabilities: the Bayesian analysis assumes that there is no uncertainty in  
438 the observed data and never considers the possibility that the observed data could be atypical in  
439 some way. If support for a clade depends critically on a single site, then the bootstrap support for  
440 that clade depends on the probability that the site will be included at least once in a particular

441 bootstrap replicate. The probability  $q$  that a particular site (out of  $n$  total sites) will be omitted  
442 from any given bootstrap data set is

$$q = \left(1 - \frac{1}{n}\right)^n, \quad (5)$$

443 which (by definition) approaches  $e^{-1}$  as  $n \rightarrow \infty$ . Thus, the probability that a single critical site  
444 will be included at least once in any given bootstrap data set is  $p = 1 - q$ , which is approximately  
445 63% for a reasonably large number of sites. We should therefore not expect strong bootstrap  
446 support for a clade if that clade is supported only by a single site, even if that site appears to be  
447 reliable indicator of history. Such a site may, however, have a strong impact on a Bayesian  
448 analysis because data sets excluding that site are never considered. For this reason, frequentist  
449 tests of data combinability that use bootstrapping to evaluate the significance of likelihood ratios  
450 are not appropriate when Bayesian approaches are used for estimating phylogeny.

## 451 SUMMARY

452 Marginal likelihoods provide a straightforward way of assessing the statistical significance  
453 of phylogenetic dissonance (Lewis et al., 2016). We simulated data sets with varying levels of deep  
454 coalescence and found, as expected, that larger numbers of deep coalescence events led to higher  
455 estimated phylogenetic dissonance and also to preference for the SEPARATE model over the  
456 CONCATENATED and HETERO models based on estimated marginal likelihoods. Exceptions  
457 mainly involved data sets with low information content due to small tree lengths, which can show  
458 low dissonance and preference for the CONCATENATED model despite a relatively large number  
459 of deep coalescence events. We calibrated BF comparisons between CONCATENATED and

460 SEPARATE using the method of García-Donato and Chen (2005) to determine the critical value  
461 that balances the prior predictive error probabilities of competing models, finding that the  
462 standard cutoff (1.0, or 0.0 on the log scale) is not always ideal but in practice changed very few  
463 of our model choice determinations. Our results also show that conflict among gene tree  
464 topologies may masquerade as heterotachy in combined analyses, as shown previously by Mendes  
465 and Hahn (2016).

#### 466 FUNDING

467 This material is based upon work supported by the National Science Foundation under  
468 grant number DEB-1354146 to POL, MHC, LK, and LAL and under grant number DEB-1036448  
469 (GrAToL) to LAL and POL. MHC's research was also partially supported by NIH grant number  
470 GM 70335 and P01 CA142538.

#### 471 ACKNOWLEDGEMENTS

472 This study benefited from computing resources made available through the computing  
473 cluster provided by the Computational Biology Core of the UConn Institute for Systems  
474 Genomics. We thank system administrator Jeffrey Lary and director Dr. Jill Wegrzyn for their  
475 assistance with these computing resources.

476

LITERATURE CITED

- 477 Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of  
478 concordance among gene trees. *Mol. Biol. Evol.* 24:412–426.
- 479 Baele, G., P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko. 2012.  
480 Improving the accuracy of demographic and molecular clock model comparison while  
481 accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* 29:2157–2167.
- 482 Bergsten, J., A. N. Nilsson, and F. Ronquist. 2013. Bayesian tests of topology hypotheses with an  
483 example from diving beetles. *Syst. Biol.* 62:660–673.
- 484 Brown, J. M. and R. C. Thomson. 2016. Bayes factors unmask highly variable information  
485 content, bias, and extreme influence in phylogenomic analyses. *Syst. Biol.* 66:517–530.
- 486 Chifman, J. and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model.  
487 *Bioinformatics* 30:3317–3324.
- 488 Chifman, J. and L. Kubatko. 2015. Identifiability of the unrooted species tree topology under the  
489 coalescent model with time-reversible substitution processes, site-specific rate variation, and  
490 invariable sites. *J. Theor. Biol.* 374:35–47.
- 491 Degnan, J. H. and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the  
492 multispecies coalescent. *Trends Ecol. Evol.* 24:332–340.
- 493 Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution*  
494 63:1–19.

- 495 Fan, Y., R. Wu, M.-H. Chen, L. Kuo, and P. O. Lewis. 2011. Choosing among partition models in  
496 Bayesian phylogenetics. *Mol. Biol. Evol.* 28:523–532.
- 497 Fučíková, K., P. O. Lewis, and L. A. Lewis. 2016. Chloroplast phylogenomic data from the green  
498 algal order Sphaeropleales (Chlorophyceae, Chlorophyta) reveal complex patterns of sequence  
499 evolution. *Mol. Phylogenet. Evol.* 98:176–183.
- 500 García-Donato, G. and M.-H. Chen. 2005. Calibrating Bayes factor under prior predictive  
501 distributions. *Statistica Sinica* 15:359–380.
- 502 Heled, J. and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data.  
503 *Mol. Biol. Evol.* 27:570–580.
- 504 Huelsenbeck, J. P. and J. Bull. 1996. A likelihood ratio test to detect conflicting phylogenetic  
505 signal. *Syst. Biol.* 45:92–98.
- 506 Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth,  
507 B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F. Zhang, H. Li,  
508 L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, M. S. Bayzid, V. Zavidovych,  
509 S. Subramanian, T. Gabaldón, S. Capella-Gutiérrez, J. Huerta-Cepas, B. Rekepalli, K. Munch,  
510 M. Schierup, B. Lindow, W. C. Warren, D. Ray, R. E. Green, M. W. Bruford, X. Zhan,  
511 A. Dixon, S. Li, N. Li, Y. Huang, E. P. Derryberry, M. F. Bertelsen, F. H. Sheldon, R. T.  
512 Brumfield, C. V. Mello, P. V. Lovell, M. Wirthlin, M. P. C. Schneider, F. Prosdocimi, J. A.  
513 Samaniego, A. M. V. Velazquez, A. Alfaro-Núñez, P. F. Campos, B. Petersen,  
514 T. Sicheritz-Ponten, A. Pas, T. Bailey, P. Scofield, M. Bunce, D. M. Lambert, Q. Zhou,  
515 P. Perelman, A. C. Driskell, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu,

- 516 J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang, J. Wang, L. Smeds, F. E. Rheindt, M. Braun,  
517 J. Fjeldsa, L. Orlando, F. K. Barker, K. A. Jønsson, W. Johnson, K.-P. Koepfli, S. O'Brien,  
518 D. Haussler, O. A. Ryder, C. Rahbek, E. Willerslev, G. R. Graves, T. C. Glenn, J. McCormack,  
519 D. Burt, H. Ellegren, P. Alström, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft,  
520 E. L. Braun, T. Warnow, W. Jun, M. T. P. Gilbert, and G. Zhang. 2014. Whole-genome  
521 analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.
- 522 Jeffreys, H. 1939. *Theory of Probability*. Oxford University Press.
- 523 Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using  
524 maximum likelihood for gene trees under coalescence. *Bioinformatics* 25:971–973.
- 525 Larget, B. 2013. The estimation of tree posterior probabilities using conditional clade probability  
526 distributions. *Syst. Biol.* 62:501–511.
- 527 Lartillot, N., T. Lepage, and S. Blanquart. 2009. PhyloBayes 3: a Bayesian software package for  
528 phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.
- 529 Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in  
530 the amino-acid replacement process. *Mol. Biol. Evol.* 21:1095–1109.
- 531 Lartillot, N. and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration.  
532 *Syst. Biol.* 55:195–207.
- 533 Leigh, J. W., E. Susko, M. Baumgartner, and A. J. Roger. 2008. Testing congruence in  
534 phylogenomic analysis. *Syst. Biol.* 57:104–115.



- 535 Lewis, P. O., M.-H. Chen, L. Kuo, L. A. Lewis, K. Fukov, S. Neupane, Y.-B. Wang, and D. Shi.  
536 2016. Estimating Bayesian phylogenetic information content. *Syst. Biol.* 65:1009–1023.
- 537 Lindley, D. V. 1956. On a measure of the information provided by an experiment. *Ann. Math.*  
538 *Stat.* 27:986–1005.
- 539 Lindley, D. V. 1957. A statistical paradox. *Biometrika* 44:187–192.
- 540 Liu, L., D. K. Pearl, R. T. Brumfield, and S. V. Edwards. 2008. Estimating species trees using  
541 multiple-allele DNA sequence data. *Evolution* 62:2080–2091.
- 542 Liu, L., S. Wu, and L. Yu. 2015. Coalescent methods for estimating species trees from  
543 phylogenomic data. *J. Syst. Evol.* 53:380–390.
- 544 Liu, L., L. Yu, and S. V. Edwards. 2010. A maximum pseudo-likelihood approach for estimating  
545 species trees under the coalescent model. *BMC Evol. Biol.* 10:302.
- 546 Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–536.
- 547 Mallet, J., N. Besansky, and M. W. Hahn. 2016. How reticulated are species? *BioEssays*  
548 38:140–149.
- 549 Mendes, F. K. and M. W. Hahn. 2016. Gene tree discordance causes apparent substitution rate  
550 variation. *Syst. Biol.* 65:711–721.
- 551 Mirarab, S., M. S. Bayzid, B. Boussau, and T. Warnow. 2014a. Statistical binning enables an  
552 accurate coalescent-based estimation of the avian tree. *Science* 346:1250463.

- 553 Mirarab, S., M. S. Bayzid, and T. Warnow. 2016. Evaluating summary methods for multilocus  
554 species tree estimation in the presence of incomplete lineage sorting. *Syst. Biol.* 65:366–380.
- 555 Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014b.  
556 ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.
- 557 Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and  
558 D. Baurain. 2011. Resolving difficult phylogenetic questions: why more sequences are not  
559 enough. *PLoS Biol.* 9:e1000602.
- 560 Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005. Phylogenomics. *Annu. Rev. Ecol.*  
561 *Evol. Syst.* 36:541–562.
- 562 Posada, D. 2016. Phylogenomics for Systematic Biology. *Syst. Biol.* 65:353–356.
- 563 Rambaut, A. and N. C. Grass. 1997. Seq-Gen: an application for the Monte Carlo simulation of  
564 DNA sequence evolution along phylogenetic trees. *Bioinformatics* 13:235–238.
- 565 Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu,  
566 M. A. Suchard, and J. P. Huelsenbeck. 2012. MrBayes 3.2: efficient Bayesian phylogenetic  
567 inference and model choice across a large model space. *Syst. Biol.* 61:539–542.
- 568 Salichos, L., A. Stamatakis, and A. Rokas. 2014. Novel Information Theory-Based Measures for  
569 Quantifying Incongruence among Phylogenetic Trees. *Mol. Biol. Evol.* 31:1261–1271.
- 570 Shannon, C. E. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.*  
571 27:379–423,623–656.

- 572 Song, S., L. Liu, S. V. Edwards, and S. Wu. 2012. Resolving conflict in eutherian mammal  
573 phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl Acad. Sci.*  
574 USA 109:14942–14947.
- 575 Swofford, D. L. 2003. PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods).  
576 Version 4. Sinauer Associates, Sunderland, Massachusetts.
- 577 Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference.  
578 Pages 407–514 *in* *Molecular Systematics*. Sinauer Associates, Sunderland, Massachusetts.
- 579 Tang, L., X. hui Zou, L. bin Zhang, and S. Ge. 2015. Multilocus species tree analyses resolve the  
580 ancient radiation of the subtribe Zizaniinae (Poaceae). *Mol. Phylogenet. Evol.* 84:232 – 239.
- 581 Venables, W. N. and B. D. Ripley. 2002. Random and mixed effects. Pages 271–300 *in* *Modern*  
582 *Applied Statistics with S* 4 ed. New York: Springer.
- 583 Wang, Y.-B., M.-H. Chen, L. Kuo, and P. O. Lewis. 2017. A new Monte Carlo method for  
584 estimating marginal likelihoods. *Bayesian Analysis* (Advance access DOI: 10.1214/17-BA1049).
- 585 Xi, Z., L. Liu, J. S. Rest, and C. C. Davis. 2014. Coalescent versus concatenation methods and  
586 the placement of Amborella as sister to water lilies. *Syst. Biol.* 63:919–932.
- 587 Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2010. Improving marginal likelihood  
588 estimation for Bayesian phylogenetic model selection. *Syst. Biol.* 60:150–160.