Running head: LIKELIHOOD AND OUTLIERS IN PHYLOGENOMICS

Title: Site and gene-wise likelihoods unmask influential outliers in phylogenomic analyses

Joseph F. Walker[1*], Joseph W. Brown[1], and Stephen A. Smith[1*]

[1]Dept. Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, 48109, USA
*Corresponding authors

Corresponding author emails: jfwalker@umich.edu, eebsmith@umich.edu

ABSTRACT

Despite the wealth of evolutionary information available from sequence data, recalcitrant nodes in phylogenomic studies remain. A recent study of vertebrate transcriptomes by Brown and Thomson (2016) revealed that less than one percent of genes can have strong enough phylogenetic signal to alter the species tree. While identifying these outliers is important, the use of Bayes factors, advocated by Brown and Thomson (2016), is a heavy computational burden for increasingly large and growing datasets. We do not find fault with the Brown and Thomson (2016) study, but instead hope to build on their suggestions and offer some alternatives. Here we suggest that site- and gene-wise likelihoods may be used to idenitfy discordant genes and nodes. We demonstrate this in the vertebrate dataset analyzed by Brown and Thomson (2016) as well as a dataset of carnivorous Caryophyllales (Eudicots: Superasterids). In both datasets, we identify genes that strongly influence species tree inference, and can overrule the signal present in all remaining genes altering the species tree topology. By using a less computationally demanding approach, we can more rapidly examine competing hypotheses, providing a more thorough assessment of overall conflict. For example, our analyses highlight that the debated vertebrate relationship of Alligatoridae sister to turtles, only has six genes with complete coverage for all species of Alligatoridae, birds and turtles. We also find that two genes (~0.0016%) from the 1237 gene dataset of carnivorous Caryophyllales drive the topological estimate and, when removed, the species tree topology supports an alternative hypothesis supported by the remaining genes. Additionally, while the genes highlighted by Brown and Thomson (2016) were revealed to be the result of errors, we suggest that the topology produced by the outlier

genes in the carnivorous Caryophyllales may not be the result of methodological error. Close examination of these genes revealed no obvious biases (i.e. no evidence of misidentified orthology, alignment error, or model violations such as significant compositional heterogeneity) suggesting the potential that these genes represent genuine, but exceptional, products of the evolutionary process. Bayes factors have been demonstrated to be helpful in addressing questions of conflict, but require significant computational effort. We suggest that maximum likelihood can also address these questions without the extensive computational burden. Furthermore, we recommend more thorough dataset exploration as this may expose limitations in a dataset to address primary hypotheses. While a dataset may contain hundreds or thousands of genes, only a small subset may be informative for the primary biological question.

INTRODUCTION

The wealth of information in phylogenomic datasets offers the potential to resolve the most difficult nodes in the tree of life. However, different datasets within the same study or published by different authors that comprise equivalent taxonomic sampling often infer competing hypotheses with high support (i.e., nonparametric bootstrap support of 100% or posterior probability of 1.0). Prominent examples of recalcitrant nodes within well-studied groups include many charismatic lineages such as the root of placental mammals (Morgan et al. 2013; Romiguier et al. 2013), early branching in Neoaves (Jarvis et al. 2014; Prum et al. 2015), and the earliest diverging lineage of angiosperms (Wickett et al. 2014; Xi et al. 2014). Understanding the evolution of these clades relies on being able to resolve these nodes. Finding the underlying causes of uncertainty in

60 phylogenomic datasets is an essential step toward resolving problematic nodes. Recently,
61 authors have developed means of exploring conflict between gene trees and species trees
62 specifically for phylogenomic datasets (Smith et al. 2015; Kobert et al. 2016; Pease et al.
63 2016), aiding in identification of regions of species trees with considerable uncertainty
64 despite strong statistical support from traditional support measures.

65 Brown and Thomson (2016) used Bayes factors as a means of uncovering genes
66 that have disproportionate influence on the reconstruction of species relationships
67 inferred from concatenated phylogenomic data. Using a previously published vertebrate
68 transcriptome dataset (Chiari et al. 2012) they found that two genes were capable of
69 altering the topology of the concatenated species tree with high support values (posterior
70 node probabilities of 1.0). Identification of these genes allowed for further analysis to
71 determine whether they were the result of errors in orthology detection or real biological
72 phenomena. While these analyses demonstrated the potential influence of strongly
73 conflicting genes on species tree construction, the reliance on a Bayes-factor approach
74 imposes an enormous computational burden on already computationally intense analyses
75 using large datasets.

76 Here we discuss an alternative to Bayes factors for identifying genes that have a
77 disproportionate contribution to the resolution of the species tree topology. We explore
78 gene-tree conflict by examining both site- and gene-specific log-likelihoods. Site-wise
79 log-likelihood analyses have been employed in phylogenomic datasets previously (Castoe
80 et al. 2009; Smith et al. 2011), primarily to compare two alternative topologies. Here, we
81 also examine per-gene log-likelihood differences and gene-wise conflict as in Smith et al.
82 (2015). We conducted these analyses on both the vertebrate dataset of Brown and

Thomson (2016) as well as a carnivorous Caryophyllales (Eudicots: Superasterids) dataset. While the two genes that were discovered by Brown and Thomson (2016) in the vertebrate dataset were identified to be errors of dataset construction, we discuss the possibility that the outlier genes found in the carnivorous dataset not be a methodological error. We hope to build upon the important conclusion drawn from Brown and Thomson (2016) by suggesting a fast and computationally feasible means of finding outlier genes in a phylogenomic dataset, discussing the importance of examining gene conflict and signal, and illustrating the possibility that some outlier genes may be the result of biological phenomena and not errors.

METHODS

Data collection

We obtained the 248 genes that were codon aligned analyzed by Brown and Thomson (2016) from Dryad deposit (http://dx.doi.org/10.5061/dryad.8gm85) of the original study (Chiari et al. 2012), focused on resolving relationships of amniotes. The coding DNA sequences of the 1237 one-to-one orthologs were downloaded from (XXXX) and used by Walker et. al (in review) to infer the relationships among carnivorous Caryophyllales (Eudicots: Superasterids). All programs and commands used in this analysis may be found at https://bitbucket.org/jfwalker/siteloglikelihood.

Species trees

Brown and Thomson (2016) used Bayesian analyses to obtain the topologies from the Chiari et. al (2012) data set. As our study focused on the use of likelihood for

106 detecting overly influential genes, we ensured that maximum likelihood would
107 recapitulate the previous species-tree results. To construct a species tree for the vertebrate
108 dataset, the 248 individual genes vertebrate genes used in Brown and Thomson (2016) for
109 inference of highly influential genes were concatenated pxcat (Brown, Walker and Smith,
110 in press). The species tree was inferred with maximum likelihood as implemented in
111 RAxML v8.2.3 (Stamatakis 2014) using the GTR+CAT model of evolution with 200
112 rapid bootstrap replicates performed. The use of CAT in the species tree analysis was
113 performed to save computational time as final inference is still conducted under
114 GAMMA. The species tree for the vertebrate dataset was inferred both with all genes
115 present, and again inferred in the same means with the previously identified two most
116 highly informative genes (8916 and 11434) removed (see below). The species tree
117 inferred through maximum likelihood and containing all data from the carnivory dataset
118 was downloaded from (XXXX). Another species tree was inferred through maximum
119 likelihood after removing the two highly informative genes (cluster575 and cluster3300;
120 see below) from the supermatrix.
121
122 *Gene tree construction and analysis of conflict*
123 Individual gene trees were inferred using maximum likelihood with the
124 GTR+CAT model of evolution as implemented in RAxML. A SH-Like test (Anisimova
125 et al. 2011), as implemented in RAxML, was performed to analyze the gene tree support.
126 As this test examines alternative topologies by NNI, it is possible that during the test a
127 topology with a higher likelihood is found. If a better topology was found during the test
128 performed for this study, that topology was used in downstream analyses. All gene trees

129 were rooted on the outgroup (*Protopterus* for the vertebrate dataset and *Beta vulgaris* and
130 *Spinacia oleraceae* for the carnivory dataset) and any gene trees not containing the
131 outgroup were left out of the conflict analysis. Conflict was assessed by examining taxon
132 bipartitions as implemented in phyparts (Smith et al. 2015) with SH-Like support of < 80
133 treated as uninformative. The conflict was mapped on the species tree using the script
134 phypartspiecharts.py (available from
135 https://github.com/mossmatters/MJPythonNotebooks).
136

137 *Gene log-likelihood Analysis*

138 The alternate topologies obtained for the placement of turtles were used along
139 with the Chiari et. al (2012) concatenated dataset for a site-wise log-likelihood analysis as
140 implemented in RAxML using the GTR+GAMMA model of evolution. The difference in
141 site-wise log-likelihoods between the two topologies, as well as the gene-wise log-
142 likelihood differences (sum of gene-specific site log-likelihoods extracted from the
143 overall matrix), were calculated using R scripts (available from
144 https://bitbucket.org/jfwalker/siteloglikelihood).
145

146 *Testing for paralogy in carnivory dataset*

147 The homolog trees created from amino acid data in the study by Walker et. al (in
148 review) were downloaded from (XXX). We examined the maximum inclusion (Yang and
149 Smith 2014) homologs of the amino acid data and compared the clusters containing the
150 outlier genes to those nucleotide clusters containing the outlier genes. This allowed us to

151 examine the possibility that the nucleotide cluster contained homology errors that would
152 be exposed by the slower evolving amino acid dataset.
153

154 ## RESULTS

155 ### *Likelihood based species tree inferences*

156 Brown and Thomson (2016) inferred a species tree with Bayesian analyses. While
157 we do not have a specific criticism of this choice, we re-analyzed the dataset using
158 maximum likelihood analyses to ensure the results were recapitulated as our study
159 focused on the use of likelihood. From the full dataset, we recovered the same topology
160 of Brown and Thomson (2016), with turtles positioned as sister to the crocodilians
161 (genera *Alligator* and *Caiman*). The edge supporting the relationship of turtles was shown
162 to have a large amount of conflict and the dominant alternative position placed turtles
163 sister to the bird clade (Fig. 1). The overall difference in log-likelihood between the two
164 topologies for the vertebrate dataset was 15.83. The removal of the vertebrate genes 8916
165 and 11434, as shown by Brown and Thomson (2016), placed turtles sister to Aves, albeit
166 with low bootstrap support (BS = 12; Supplementary Fig. 1). In the carnivorous
167 Caryophyllales, the inferred species tree contained two edges with many conflicting gene
168 trees and one dominant alternative topology (Fig. 1). The edge supporting *Ancistrocladus*
169 and *Drosophyllum* sister to the rest of carnivorous plants received no bootstrap support
170 (BS = 0); however, when reanalyzed with cluster575 and cluster3300 removed the
171 position of *Ancistrocladus* and *Drosophyllum* changed and the relationship gained
172 support (BS = 100; Supplementary Fig. 1). The log-likelihood difference between the
173 Caryophyllales topologies was 74.94.

174 *Gene tree conflict and log-likelihood analysis shows genes of disproportionate influence*

175 For the vertebrate dataset, we limited conflict analysis to genes that contained
176 sequences for the outgroup *Protopterus*. This resulted in 93 (of 248) usable gene trees for
177 conflict analysis. Many genes were missing one or more taxa, with only five genes
178 containing information for all ingroup taxa (Table 1). Throughout the vertebrate tree, we
179 found conflict at many of the deeper nodes (Fig. 1). Also, the node representing the
180 controversial placement of turtles as sister to crocodilians had only seven genes with high
181 SH support (>80). Nine genes recovered, with high SH support, supported the dominant
182 alternative relationship of turtle's sister to a clade comprising of Alligatoridae and birds
183 with high SH support.

184 The site-wise log-likelihood analyses did not clearly identify major biases (Fig.
185 2A and Fig. 2C). The gene-wise log-likelihood comparison of the two dominant
186 topologies showed that two genes (ENSGALG00000008916 and
187 ENSGALG00000011434) exhibit a disproportionate influence on the overall likelihood
188 of the supermatrix (Fig. 2B). The genes identified using the likelihood approach
189 presented here were the same genes identified by Brown and Thomson (2016) using
190 Bayes factors. The genes had a difference in log-likelihood scores of 72.91 and 41.71,
191 respectively, and support the hypothesis of turtles sister to crocodilians with an average
192 difference in log-likelihood of any gene in the supermatrix being 3.28.

193 We performed this same comparison of log-likelihoods between the dominant
194 topologies on the carnivory dataset. We found two genes (cluster575 and cluster3300)
195 that contribute disproportionately to the overall likelihood and that individually have a

196 difference in log-likelihood scores of 33.06 and 16.63, respectively, with the average
197 difference of log-likelihood of a gene in the supermatrix to either topology being 2.882.
198
199 *Disproportionate information may potentially be a biological reality*
200 For the carnivorous Caryophyllales dataset, we explored the possibility that the
201 strongly conflicting genes cluster575 and cluster3300 reflected some methodological
202 error in the assembly pipeline, as is the case for the genes identified by Brown and
203 Thomson (2016). However, both the alignment and inferred phylogram for each gene
204 revealed no obvious problems or potential sources of systematic error (sparse alignment,
205 abnormally long branch lengths etc…). We also explored whether compositional
206 heterogeneity could explain the strongly conflicting results (i.e., that the relationships
207 were not truly conflicting, but instead incorrectly modeled). However, both RY-coding in
208 RAxML and explicit modeling of multiple equilibrium frequencies (2, 3, or 4
209 composition regimes) across the tree in p4 v1.0 (Foster 2004) failed to overturn the
210 inferred relationships. We further explored the possibility of misidentified orthology. By
211 examining the homolog tree produced from amino acid data, we identified the ortholog
212 from the nucleotide data to be complete (i.e., an ortholog within the homolog amino acid
213 tree). We found that with the slower amino acid data the sequences in the nucleotide
214 cluster575 were inferred as a single monophyletic ortholog within a duplicated homolog
215 (Supplementary Figure 2). The discrepancies that appeared between the amino acid
216 dataset and the CDS dataset were found to be either different in-paralogs/splice sites
217 maintained during the dataset cleaning procedure or short sequences that were not
218 identified as homologs in the CDS dataset (Supplementary table and Supplementary

Figure 2).

DISCUSSION

We found few genes that strongly supported the deeper relationships in the vertebrate dataset (Fig. 1). Biological processes including substitution saturation, hybridization, horizontal gene transfer, and incomplete lineage sorting can contribute to conflicting signal and may explain both the conflict and lack of information. However, limitations in the dataset might also be a factor, as few gene regions contained sequence data for every species (Table 1). In fact, only six genes have complete sampling of all species (birds, turtles and Alligatoridae) involved in examining the position of turtles in relation to the crocodilian clade. Thirty-six genes had the species sampling necessary to address the alternate hypothesis, with turtle's sister to birds. Surprisingly, only five of the genes in the analysis contained information for all ingroup taxa (Table 1). Despite the size of many of these phylogenomic datasets, the available data to address specific questions may be significantly smaller and should be analyzed as it has been shown that taxon sampling influences even large phylogenomic datasets (Walker et. al in review).

As has been noted by several authors, gene tree conflict and concordance should be examined within phylogenomic datasets (Salichos et al. 2014; Smith et al. 2015; Kobert et al. 2016). High support values can mask significant underlying conflict (Ryan et al. 2013; Salichos et al. 2014; Wickett et al. 2014; Smith et al. 2015; Yang et al. 2015; Kobert et al. 2016). This is clearly the case for the vertebrates (Chiari et al. 2012; Crawford et al. 2015; Brown and Thomson 2016). Although the vertebrate dataset contained gene tree conflict, significant missing data, and small likelihood differences

242 among alternate topologies, high posterior probabilities were reported at every node (PP
243 = 1.00) (Brown and Thomson 2016). The change in topology of the carnivory dataset is
244 also remarkable, as all 1237 genes are represented in each species and the removal of two
245 (0.0016%) of the genes resulted in a different topology with high support. Some authors
246 have noted discrepancies between coalescent and supermatrix results in these datasets
247 (Wickett et al. 2014; Xi et al. 2014;Walker et. al in review). The observation that a small
248 number of genes, in the context of supermatrix analyses, can influence the resulting
249 topology may help explain this phenomenon. The results from both datasets discussed
250 here emphasize that even with large, high coverage datasets, supermatrix analyses may be
251 sensitive to a small number of influential genes.

252 We found relatively small differences in overall log-likelihood scores between the
253 alternate competing hypotheses for both the carnivore and vertebrate datasets. Through a
254 site- and gene-wise log-likelihood comparison, we demonstrated the presence of genes
255 that disproportionally contribute to the species-tree inference and examine the strong
256 conflicting signal. While Bayes factors provide another means of finding these genes,
257 they are computationally expensive, which is a major concern given the growing size of
258 phylogenomic datasets. Using the site- and gene-wise likelihood approach, we identified
259 the genes that have a disproportionate effect on the likelihood in ~400 seconds using two
260 processors on a laptop. Identifying these genes is important for understanding potential
261 errors, biological processes (hybridization, horizontal gene transfer), and to avoid
262 violating model assumptions where strong conflicting signal cannot be incorporated.
263 Identifying these genes *quickly* allows for more thorough examination of the entire
264 dataset in addition to the outliers (not to mention the CPU years and carbon savings).

The two outlier genes in the vertebrate dataset were demonstrated to be misidentified orthologs (Brown and Thomson 2016). Unfortunately, the genomic resources are not available to fully examine the carnivorous outlier genes (e.g., we do not have synteny or information on gene loss). However, we used tools and data such as alignment analysis, compositional heterogeneity tests, and homolog analysis to examine the two carnivorous genes to the best of our ability. Our analyses did not detect any problems with the alignment or composition. Our homolog analyses identified one gene, cluster 575, to be an ortholog of a gene that experienced a duplication. While we cannot rule out every possible source of error, we also cannot identify a source of methodological error, suggesting the possibility that the conflicting topology is the result of real (albeit unknown) biological processes.

CONCLUSION

Brown and Thomson (2016) used Bayes factors to identify the phylogenetic signal in genes and discovered that two genes supporting a topology conflicting with the dominant topology can dominate species-tree inference. Although Bayes factors are a powerful method of identifying support for topological relationships in a Bayesian context, they are computationally expensive. We show that likelihood analyses, which are significantly less computationally intensive, can also identify these genes in phylogenomic datasets. This lower computational burden frees resources for more thorough analyses of conflict of the identified outlier genes. We show that, despite the size of many of these datasets, very few genes can address key topological questions due to missing data and/or saturation. Additionally, the genes identified by Brown and

Thomson (2016) were suggested to be the result of error in orthology detection. In the carnivory dataset, we also identify genes with strong conflicting signal, but this might be the result of biological processes and not methodological error. The paper by Brown and Thomson (2016) lays an exciting framework for exploring data used in phylogenomic analyses, and we further highlight the importance of their finding. We show that for a dataset of 1237 genes, removing two genes (0.0016%) alters the topology and provides high support. Collectively these finding show that the potential impact of a small number of genes on the estimation of species trees is a critical topic for further examination.

FUNDING

JFW is supported through a fellowship provided by University of Michigan department of Ecology and Evolutionary Biology. JFW and SAS were supported by NFS 1354048 and JWB and SAS were supported by NSF 1207915.

ACKNOWLEDGEMENTS

We would like to thank Oscar Vargas, Greg Stull and Ning Wang for discussion of the manuscript and methods.

REFERENCES

Anisimova M., Gil M., Dufayard J.F., Dessimoz C., Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst. Biol. 60:685–699.

311 Brown J.M., Thomson R.C. 2016. Bayes factors unmask highly variable information
312 content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. syw101.
313 Brown J.W., Walker J.F., Smith S.A. 2017. phyx: Phylogenetic tools for Unix.
314 Bioinformatics. accepted.
315 Castoe T.A., de Koning A.P.J., Kim H.-M., Gu W., Noonan B.P., Naylor G., Jiang Z.J.,
316 Parkinson C.L., Pollock D.D. 2009. Evidence for an ancient adaptive episode of
317 convergent molecular evolution. Proc. Natl. Acad. Sci. 106:8986–8991.
318 Chiari Y., Cahais V., Galtier N., Delsuc F. 2012. Phylogenomic analyses support the
319 position of turtles as the sister group of birds and crocodiles (Archosauria). BMC
320 Biol. 10:65.
321 Crawford N.G., Parham J.F., Sellas A.B., Faircloth B.C., Glenn T.C., Papenfuss T.J.,
322 Henderson J.B., Hansen M.H., Simison W.B. 2015. A phylogenomic analysis of
323 turtles. Mol. Phylogenet. Evol. 83:250–257.
324 Foster P.G. 2004. Modeling compositional heterogeneity. Syst Biol. 53:485–495.
325 Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C.,
326 Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li
327 H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S.,
328 Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J.,
329 Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E.,
330 Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P.,
331 Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C. V., Lovell P. V., Wirthlin
332 M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M. V., Alfaro-
333 Nunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield

334 P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B.,
335 Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang
336 Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L.,
337 Barker F.K., Jonsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D.,
338 Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J.,
339 Burt D., Ellegren H., Alstrom P., Edwards S. V., Stamatakis A., Mindell D.P.,
340 Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-
341 genome analyses resolve early branches in the tree of life of modern birds. Science
342 (80-.). 346:1320–1331.
343 Kobert K., Salichos L., Rokas A., Stamatakis A. 2016. Computing the internode certainty
344 and related measures from partial gene trees. Mol. Biol. Evol. Advance Ac:1–17.
345 Morgan C.C., Foster P.G., Webb A.E., Pisani D., McInerney J.O., O'Connell M.J. 2013.
346 Heterogeneous models place the root of the placental mammal phylogeny. Mol. Biol.
347 Evol. 30:2145–56.
348 Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics Reveals Three
349 Sources of Adaptive Variation during a Rapid Radiation. PLoS Biol. 14:1–24.
350 Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Moriarty Lemmon E.,
351 Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted
352 next-generation DNA sequencing. Nature. 526:569–573.
353 Romiguier J., Ranwez V., Delsuc F., Galtier N., Douzery E.J.P. 2013. Less is more in
354 mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the
355 root of placental mammals. Mol. Biol. Evol. 30:2134–44.
356 Ryan J.F., Pang K., Schnitzler C.E., Nguyen A.D., Moreland R.T., Simmons D.K., Koch

357 B.J., Francis W.R., Havlak P., Smith S.A., Putnam N.H., Haddock S.H., Dunn C.W.,
358 Wolfsberg T.G., Mullikin J.C., Martindale M.Q., Baxevanis A.D. 2013. The genome
359 of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution.
360 Science (80-.). 342:1242592.
361 Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for
362 quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.
363 Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets
364 reveals conflict, concordance, and gene duplications with examples from animals
365 and plants. BMC Evol. Biol. 15:150.
366 Smith S.A., Wilson N.G., Goetz F.E., Feehery C., Andrade S.C.S., Rouse G.W., Giribet
367 G., Dunn C.W. 2011. Resolving the evolutionary relationships of molluscs with
368 phylogenomic tools. Nature. 480:364–367.
369 Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-
370 analysis of large phylogenies. Bioinformatics. 30:1312–1313.
371 Walker, J.F., Yang, Y. Moore, M.J., Mikenas, J., Brockington, S.F., Timoneda, A., Smith
372 S.A. 2017. Widespread paleopolyploidy, gene tree conflict and recalcitrant
373 relationships among the carnivorous caryophyllales. in review.
374 Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N.,
375 Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R.,
376 Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis
377 P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson
378 D.W., Surek B., Villarreal J.C., Roure B., Philippe H., DePamphilis C.W., Chen T.,
379 Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y.,

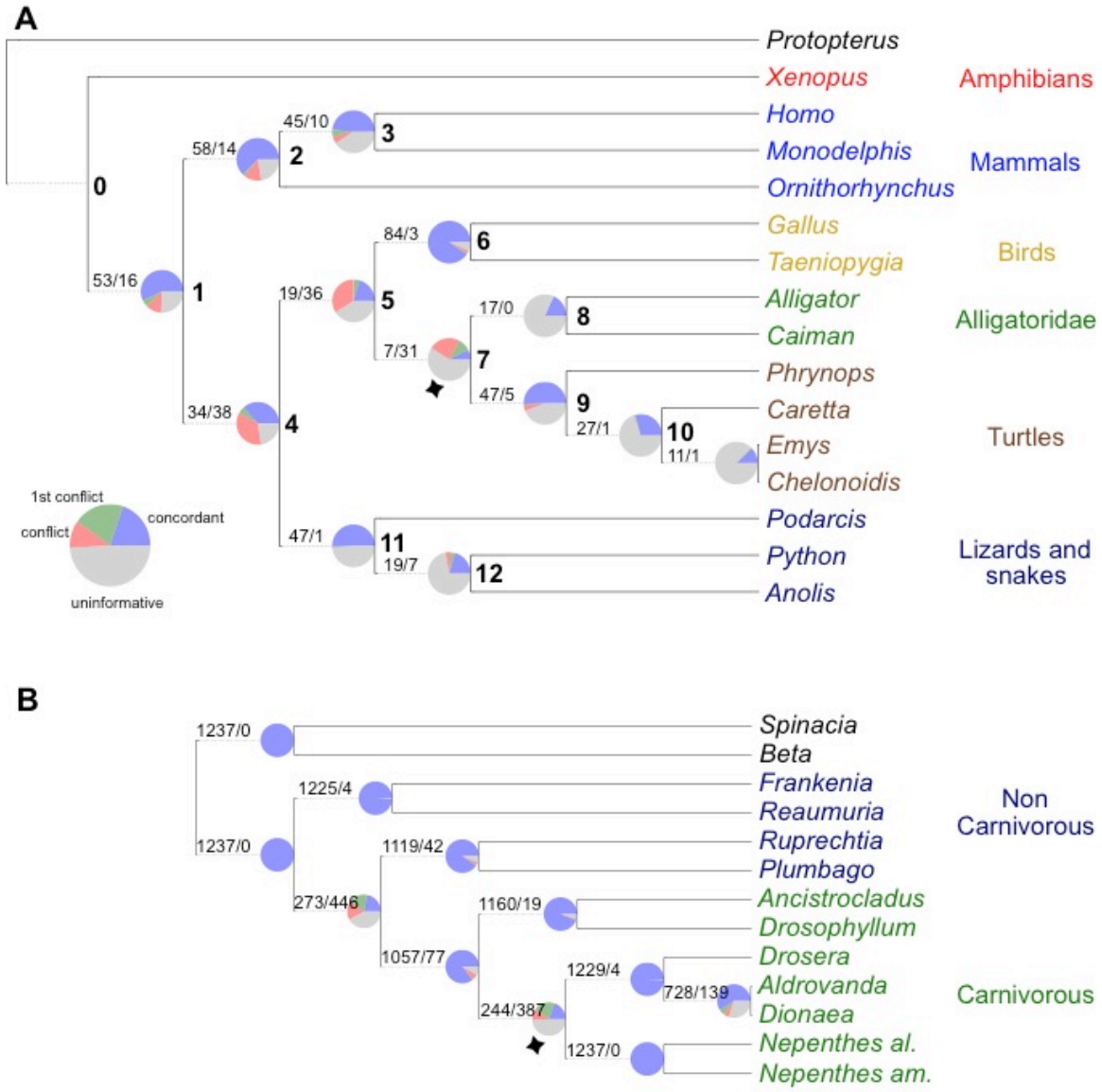
380 Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S., Leebens-Mack J. 2014.
381 Phylotranscriptomic analysis of the origin and early diversification of land plants.
382 Proc. Natl. Acad. Sci. 111:E4859–E4868.
383 Xi Z., Liu L., Rest J.S., Davis C.C. 2014. Coalescent versus Concatenation Methods and
384 the Placement of Amborella as Sister to Water Lilies. Syst. Biol. 63:919–932.
385 Yang Y., Moore M.J., Brockington S.F., Soltis D.E., Wong G.K., Carpenter E.J., Zhang
386 Y., Chen L., Yan Z., Xie Y., Sage R.F., Covshoff S., Hibberd J.M., Nelson M.N.,
387 Smith S.A. 2015. Dissecting molecular evolution in the highly diverse plant clade
388 Caryophyllales using transcriptome sequencing. Mol. Biol. Evol.:msv081-.
389
390
391
392
393
394
395
396
397
398
399
400
401
402

403 FIGURE CAPTIONS

404

405 **Figure 1. Maximum likelihood trees inferred by RAxML for the two datasets.**
406 Conflict analysis for the vertebrates (A) and carnivorous Caryophyllales (B). The
407 vertebrate analysis includes the 93 genes that contained the outgroup (*Protopterus*), and
408 the carnivorous Caryophyllales includes 1237 genes all of which had the outgroups
409 (*Spinacia oleraceae* and *Beta vulgaris*). Grey represents uninformative genes (SH-Like <
410 80 or no taxon representation for the bipartition), blue represents gene trees that are
411 concordant with the relationship, green represents the dominant alternate topology and
412 red represents all other conflict. Numbers on edges represent concordance/conflict.
413 Numbers at the nodes of the vertebrate dataset correspond to bipartition numbers in Table
414 1.

415



416

417

418

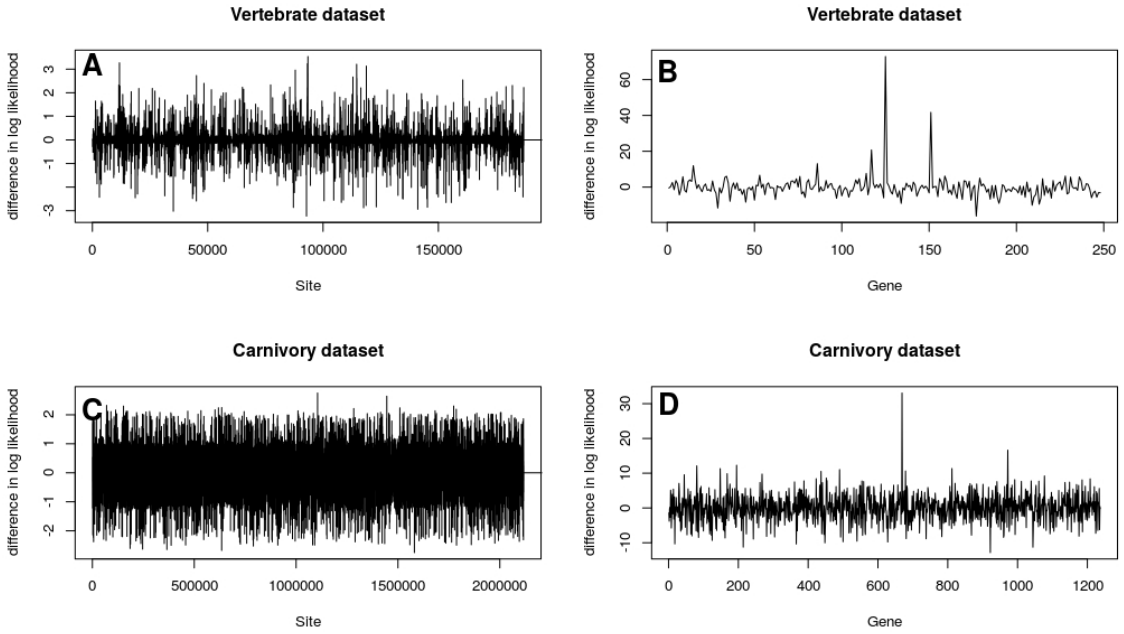
419

420

421

422

423 **Figure 2. Site-specific log-likelihood analysis between competing hypotheses**. Plots A
424 and C show the difference in log-likelihood for each site and plots B and D show the
425 difference in log-likelihood for each gene. Positive values in plots A and B indicate
426 support for turtles sister to crocodilians and negative values indicate support for turtles
427 sister to a clade consisting of birds and Alligatoridae. Positive values in plots C and D
428 indicate support for *Drosophyllum lusitanicum* and *Ancistrocladus robertsonorium* sister
429 to other carnivorous plants and negative values indicate support for *D. lusitanicum* and *A.*
430 *robertsonorium* forming a clade with the *Nepenthes* samples.
431

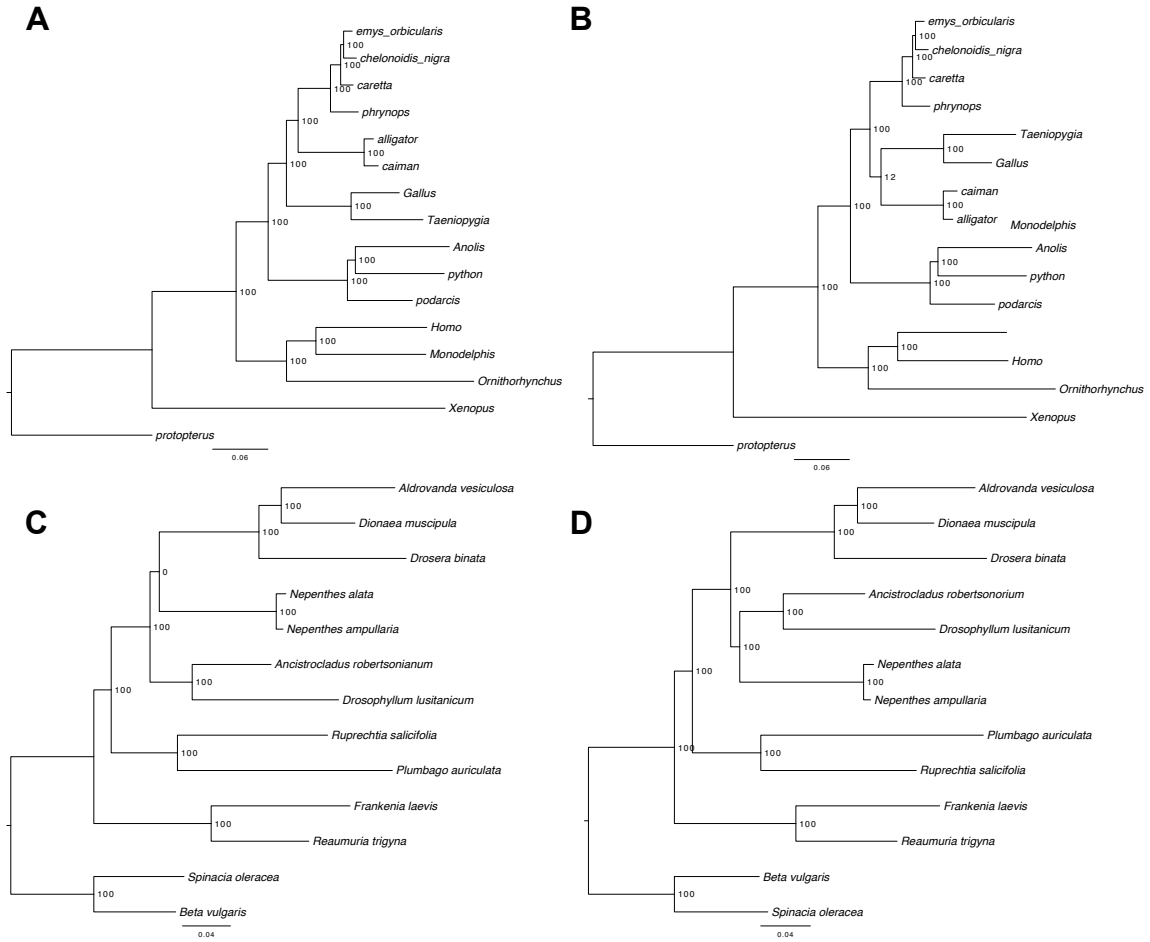


432
433
434
435
436

Table 1. Number of gene trees in which all the species for a given bipartition are present. Bipartitions correspond to node labels on figure 1.

Bipartition number	Gene trees containing all species in the bipartition
0	5
1	5
2	246
3	248
4	5
5 (All turtles, Alligatoridae and birds)	6
6	248
7	6
8	23
9	36
10	45
11	69
12	51
13	94
Bipartition of turtles sister to a clade of birds and Alligatoridae	36

Supplementary Figure 1. Species trees inferred using maximum likelihood from the different supermatrices. Support at each node was obtained from 200 rapid bootstrap replicates. A) Carnivorous Caryophyllales species tree inferred from all 1237 genes. B) Carnivorous Caryophyllales species tree inferred with cluster575 and cluster3300 removed from the supermatrix. C) Species tree for vertebrate dataset inferred with all 248 genes included in the supermatrix. D) Species tree for the vertebrate dataset inferred with 8916 and 11434 removed from the supermatrix.



452 **Supplementary Figure 2. Homolog tree for Amino Acid clustered (726) and CDS**

453 **clustered (575) highly influential gene in the carnivorous Caryophyllales dataset.**

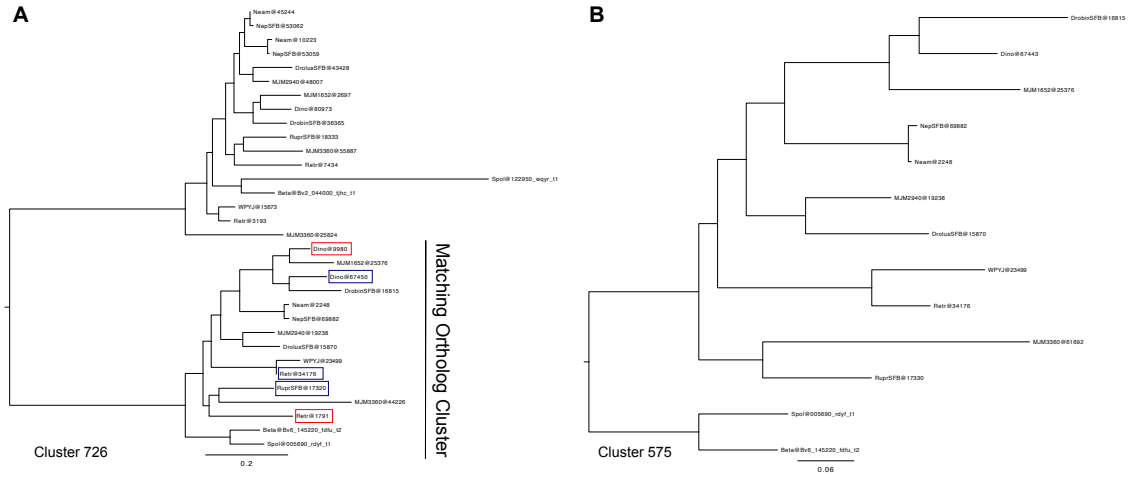
454 Different genes identified in the ortholog clusters are circled on cluster 726. Genes

455 circled in red represent ones that are shorter and were not identified as orthologous in the

456 CDS dataset and genes circled in blue represent alternate paralogs or introsplice sites

457 used between the two clustering analyses.

458



459

460

461

462

463

464

465

466

467

468 **Supplementary Table. Sources of discrepancy between the orthologs detected in**

469 **highly influential nucleotide cluster575 and in matching amino acid homolog**

470 **cluster726.**

Ortholog in 575	Ortholog in 726	Seq length of 575 (Nuc)	Seq length of 726 (Nuc)	Reason for misidentification
Dino@67443 (*Dionaea*)	Dino@67450	2793	2991	Different copy of the in-paralog or intron splice site was retained
Dino@67443 (*Dionaea*)	Dino@9980	2793	510	Not identified as homologs in blast
RuprSFB@17320 (*Ruprechtia*)	RuprSFB@17330	2787	2787	Different copy of the in-paralog or intron splice site was retained
MJM3360@61692 (*Plumbago*)	MJM3360@44226	2211	2403	Different copy of the in-paralog or intron splice site was retained
Retr@34176 (*Reaumuria*)	Retr@1791	1044	546	Not identified as homologs in blast

471

472