Running head: LIKELIHOOD AND OUTLIERS IN PHYLOGENOMICS

Title: Analyzing contentious relationships and outlier genes in phylogenomics

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

[1]Deptartment 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 genomic and transcriptomic data, recalcitrant relationships in phylogenomic studies remain throughout the tree of life. Recent studies have demonstrated that conflict is common among gene trees, and less than one percent of genes may ultimately drive species tree inference. In this study, we examined a plant and a vertebrate dataset where supermatrix and coalescent-based species trees conflict. Using a two-topology site-specific log-likelihood test, we identified two highly influential genes in each dataset. While the outlier genes in the vertebrate dataset have been shown to be the result of errors in orthology detection, we demonstrate that the outlier genes from the plant dataset may be the result of biological processes rather than model or methodological errors. When the outlier genes were removed from each supermatrix, the inferred trees matched the topologies obtained from coalescent analyses. While most tests of this nature limit the comparison to two trees, gene tree topologies generated under processes such as incomplete lineage sorting are unlikely to precisely match either topology in a two-topology test. We therefore examined bipartitions across a set of trees and recover more support for the resolution favored by coalescent analyses. These results suggest that by expanding beyond two-topology comparisons, we can dramatically improve our understanding of the underlying signal in phylogenomic datasets by asking more targeted edge-based questions.

INTRODUCTION

Many recent studies have highlighted that small changes to a dataset can yield conflicting hypotheses at particular recalcitrant relationships with high support (i.e., 100% support from nonparametric bootstrap (BS) or posterior probability (PP) values). Prominent examples of this 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 extant angiosperms (Zanis et al. 2002; Wickett et al. 2014; Xi et al. 2014). The resolution of these relationships is critical to understanding the evolutionary history of their respective clades (e.g., patterns of biochemical, morphological, and life history evolution).

Finding the underlying causes of uncertainty in phylogenomic datasets is an essential step toward resolving problematic relationships. Recently, authors have developed means of exploring conflict between gene trees and species trees specifically for phylogenomic datasets (Salichos et al. 2014; Smith et al. 2015; Kobert et al. 2016), aiding in the identification of regions of species trees with considerable uncertainty despite strong statistical support from traditional support measures. Two studies have shown that the disproportionate influence of just one or two genes on a supermatrix analysis is capable of altering tree topology inference (Brown and Thomson 2017; Shen et al. 2017). Brown and Thomson (2017) reanalyzed a series of datasets using Bayes factors and found that the transcriptome data from Chiari et al. (2012) contained outlier genes. When the outlier genes were included in phylogenetic reconstruction, a clade of turtles+crocodilians was inferred to be sister to birds with 100% PP. The same topology was previously inferred using ML in the original study by Chiari et al. (2012), but was

62 dismissed in favor of a coalescent reconstruction that placed turtles sister to
63 birds+crocodilians. When Brown and Thomson (2017) removed the two outlier genes,
64 the reduced supermatrix infers the same topology as the coalescent reconstruction with
65 100% PP. Another recently published study used a two-topology gene-wise likelihood
66 approach to examine contentious relationships across the tree of life and found
67 disproportionate influence of genes at all relationships examined (Shen et al. 2017).
68 Given the prevalence of outlier genes in phylogenomic datasets, and the continued
69 focus on contentious relationships in the tree of life, it is imperative that we develop
70 methods for analyzing conflict and selecting among alternative resolutions for recalcitrant
71 relationships. We build upon the discussions of Brown and Thomson (2017) and Shen et
72 al. (2017) by addressing whether these outlier genes violate models of evolution and
73 expanding beyond a comparison of only two-topologies by examining bipartitions (i.e.,
74 edges defining reciprocally monophyletic clades), across sites and genes. Typically, site-
75 wise and gene-wise log-likelihood analyses of phylogenomic datasets compare two
76 alternate topologies (e.g., Castoe et al. 2009; Smith et al. 2011; Shen et al. 2017).
77 However, given widespread gene tree discordance (e.g., due to incomplete lineage sorting,
78 ILS), it may be more realistic to assume that many alternative topologies are supported
79 within larger genomic datasets (e.g., Smith et al. 2015; Pease et al. 2016; Walker et al.
80 2017). Additionally, when the research question involves a single relationship and not the
81 entirety of the tree, it may be more appropriate to examine bipartitions instead completely
82 resolved topologies. This allows for any processes that may be causing conflict in the
83 non-focal parts of the tree to be accommodated without influencing the relationships of
84 interest. Here, we compare results from two-topology gene-wise log-likelihood analyses,

maximum gene-wise bipartition (MGWB) analyses, and maximum site-wise bipartition (MSWB) analyses. We applied these measures to vertebrate (Chiari et al. 2012; Brown and Thomson 2017) and carnivorous Caryophyllales datasets (Walker et al. 2017) (the latter hereafter referred to as the carnivory dataset). Both datasets contain contentious relationships, outlier genes, and, in their respective original studies, the authors dismissed the supermatrix topology for the topology inferred using a coalescent method.

METHODS

Data collection

We obtained the 248 genes that were codon-aligned and analyzed by Brown and Thomson (2017) from Dryad deposit (http://dx.doi.org/10.5061/dryad.8gm85) of the original study (Chiari et al. 2012) that focused on resolving the placement of turtles among amniotes. The coding DNA sequences of the 1237 one-to-one orthologs from Walker et al. (2017) to infer the relationships among carnivorous Caryophyllales (Eudicots: Superasterids) are available from Dryad (http://datadryad.org/resource/doi:10.5061/dryad.vn730). All programs used in this analysis may be found at https://bitbucket.org/jfwalker/maximizelikelihoods.

Species trees

Brown and Thomson (2017) used Bayesian analyses to obtain the topologies from the Chiari et al. (2012) data set. As our study focused on the use of maximum likelihood (ML) for detecting overly influential genes, we ensured that ML phylogenetic reconstruction would recapitulate the previous species-tree results. To construct a

108 supermatrix tree for the vertebrate dataset, the 248 individual vertebrate genes used in
109 Brown and Thomson (2017) were concatenated using the Phyx program pxcat (Brown et
110 al. 2017). The species tree was inferred in RAxML v8.2.3 (Stamatakis 2014) using the
111 GTR+Γ model of evolution, and edge support was assessed from 200 rapid bootstrap
112 replicates. Supermatrix trees for the vertebrate dataset were inferred both with all genes
113 present, and again with the previously identified two most highly informative genes
114 (8916 and 11434) removed (see below). The ML inferred from all the data from the
115 carnivory dataset was downloaded from (http://dx.doi.org/10.5061/dryad.33m48) while a
116 novel ML tree was inferred from a reduced supermatrix that excluded two highly
117 informative genes (cluster575 and cluster3300; see below).
118
119 *Gene tree construction and analysis of conflict*
120 Individual gene trees for both datasets were inferred using ML with the GTR+ Γ model of
121 evolution as implemented in RAxML. A SH-like test (Anisimova et al. 2011), as
122 implemented in RAxML, was performed to assess gene tree edge support. As this test
123 examines alternative topologies by nearest-neighbor interchange (NNI), it is possible that
124 during the test a topology with a higher likelihood is found (i.e., an 'NNI-optimal'
125 topology). When a better topology was found during the test performed for this study,
126 that topology was used in downstream analyses. We used the pxrr program in the Phyx
127 package (Brown et al. 2017) to root all gene trees on the outgroup (*Protopterus* for the
128 vertebrate dataset, and *Beta vulgaris* and *Spinacia oleraceae* for the carnivory dataset)
129 and we excluded gene trees where an outgroup was not present. We mapped conflict onto
130 the supermatrix tree using phyparts (Smith et al. 2015) with SH-like support of < 80

131 treated as uninformative. Gene tree conflict was visualized using the script
132 phypartspiecharts.py (available from
133 https://github.com/mossmatters/MJPythonNotebooks. We conducted more detailed
134 conflict analyses used for bipartition comparisons discussed below using pxbp as part of
135 the Phyx package (Brown et al. 2017).
136
137 *Calculating two-topology gene-wise log-likelihoods*
138 The alternate topologies (supermatrix and coalescent) and data matrices for the vertebrate
139 and carnivory datasets were obtained from the original studies, Chiari et al. (2012) and
140 Walker et al. (2017), respectively. We calculated site-wise log-likelihood scores for the
141 two topologies in RAxML using the GTR+ Γ model of evolution, with the data
142 partitioned by gene. The differences in site-wise log-likelihoods between the candidate
143 topologies were then calculated using scripts available from
144 https://bitbucket.org/jfwalker/maximizelikelihoods.
145
146 *Calculating maximum gene-wise and site-wise bipartitions*
147 We expanded beyond the fixed two-topology comparison by examining gene- and site-
148 wise log likelihoods over a set of trees and estimating the maximum gene-wise bipartition
149 (MGWB) and maximum site-wise bipartition (MSWB). The candidate set of trees
150 consisted of all unique gene trees as well as trees inferred from ML supermatrix and
151 coalescent analyses. For each tree in the candidate set, we calculated gene- and site-wise
152 likelihoods using RAxML with the GTR+Γ model of evolution. We then determined the
153 set of bipartitions in each dataset that conflicted with the supermatrix reconstruction of

154 the focal edge and identified which trees from the tree set resolved to which bipartition.
155 We calculated which tree representing each bipartition had the highest likelihood for each
156 gene and site. The MGWB was then identified as the bipartition containing the tree with
157 the highest log-likelihood. Because multiple trees can resolve the same bipartition for the
158 relationship of interest, the same tree need not have the highest likelihood for each gene
159 in order for the same bipartition to be the MGWB for each gene (e.g., ((A,B),C),(D,E))
160 and ((A,C),B),(D,E)) both have the bipartition separating A,B,C from D,E). The MSWBs
161 were found in the same way as the MGWBs but by site-wise log-likelihood instead of
162 gene-wise. We then recorded the frequency that each bipartition was identified as a
163 MGWB and MSWB. While we record the maximum log-likelihood for each site for each
164 bipartition, these are not comparable among bipartitions directly as the count of trees
165 displaying each bipartition varies and hence they differ in the number of estimated
166 parameters involved. However, as a coarse comparison we summed the gene-wise log-
167 likelihoods for the three bipartitions most frequently recognized as MGWBs and then
168 calculated AIC scores for each bipartition, accommodating for the increase in model
169 complexity (i.e., the number of topologies used for each bipartition).
170
171 *Testing for paralogy in carnivory dataset*
172 The homolog trees created from amino acid data in the study by Walker et al. (2017)
173 were downloaded from Dryad (http://datadryad.org/resource/doi:10.5061/dryad.vn730).
174 We examined the homologs clustered using amino acids as opposed to nucleotides. This
175 allowed us to examine whether a nucleotide cluster contained homology errors that may

176 be exposed by the slower evolving amino acid dataset. The code needed to perform this
177 analysis may be found at (https://bitbucket.org/jfwalker/maximizelikelihoods).
178

179 ## RESULTS

180 *Gene tree conflict and log-likelihood analysis reveals genes of disproportionate influence*

181 Our analysis of the vertebrate dataset recovered the same supermatrix topology (Fig. 1) as
182 found with ML by Chiari et al. (2012) and Bayesian inference by Brown and Thomson
183 (2017). The difference in log-likelihood between the supermatrix and coalescent
184 topologies for the vertebrate dataset was 15.83. Ninety-three of 248 gene trees could be
185 rooted on the outgroup *Protopterus* and only five of these had all taxa represented
186 (Supplementary Table 1). We found low support for relationships within gene trees (SH
187 <80) and significant gene tree conflict (Fig. 1). Of the gene trees with high support (SH
188 >80), seven resolved turtles+crocodilians as sister to birds (hereafter referred to as the
189 vertebrate supermatrix topology) and nine resolved crocodilians+birds sister to turtles
190 (hereafter referred to as the vertebrate coalescent topology).

191 The two-topology gene-wise log-likelihood comparison showed that 107 genes
192 had a higher likelihood score for the vertebrate supermatrix topology while 141 supported
193 the vertebrate coalescent topology (Figs. 2A, 3A). Two genes (ENSGALG00000008916
194 and ENSGALG00000011434, referred to here as 8916 and 11434, respectively) exhibited
195 a disproportionate influence on the overall likelihood of the supermatrix (Fig. 2A). These
196 two genes identified with maximum likelihood analyses matched those previously
197 identified as outliers using Bayes factors (Brown and Thomson 2017). These two genes
198 both supported the vertebrate supermatrix topology with log-likelihood scores of 72.91

199 and 41.71 greater than the alternative coalescent tree topology, respectively. The
200 difference in log-likelihood between the two topologies of the non-outlier genes ranged
201 from |0.015| to |20.756| with an average of |3.28| for all genes in the analysis. The
202 removal of the vertebrate genes 8916 and 11434, as shown by Brown and Thomson
203 (2017), recovered the coalescent topology, albeit with low bootstrap support (BS = 12;
204 Supplementary Fig. 1).

205 Previous work on the carnivory dataset, demonstrated that the placement of the
206 *Ancistrocladus+Drosophyllum* clade (Fig. 1) contained significant conflict and is
207 strongly influenced by species sampling (Walker et al. 2017). The log-likelihood
208 difference between the supermatrix and coalescent topologies was 74.94. The two-
209 topology log-likelihood comparison between the dominant topologies on the carnivory
210 dataset (Fig. 2C) showed that 623 genes supported *Ancistrocladus+Drosophyllum* sister
211 to all other carnivorous plants (hereafter referred to as carnivory supermatrix topology)
212 while 614 genes supported *Ancistrocladus+Drosophyllum* sister to *Nepenthes*
213 *alata+Nepenthes ampullaria* (hereafter referred to as carnivory coalescent topology; Fig.
214 2C&3D). Two genes (cluster575 and cluster3300) contributed disproportionately to the
215 overall likelihood. Individually these two genes have a difference in log-likelihood scores
216 between the two topologies of 33.06 and 16.63, respectively, and support the carnivory
217 supermatrix topology. When we reanalyzed the supermatrix with cluster575 and
218 cluster3300 removed, the carnivory coalescent topology was recovered, with 100% BS
219 support (Supplementary Fig. 1). The difference between the two topologies in log-
220 likelihood of the non-outlier genes ranged from |0.005| to |12.296| with an average of
221 |2.88| for all genes in the analysis.

222

223 *Maximum Gene-wise and Site-wise bipartitions change supported topology*

224 We compared MGWB and MSWB involving the contentious bird, crocodilian, and turtle
225 relationships in the vertebrate dataset (Figs. 2B, 3B,C). We found seven unique
226 topologies with the necessary species coverage to conduct the analyses: five gene tree
227 topologies from Chiari et al. (2012) and the two dominant species tree topologies. The set
228 of seven trees included three major conflicting bipartitions for the relationship in
229 question: the two resolutions found in the supermatrix and coalescent trees and birds
230 sister to crocodilian+mammals+turtles. 88 genes supported the vertebrate supermatrix
231 bipartition, 150 genes supported the vertebrate coalescent bipartition, and 10 genes
232 supported the third conflicting bipartition (Figs. 2B, 3B). The AIC scores for the summed
233 likelihoods of the supermatrix (three source trees), the coalescent (three source trees), and
234 the third bipartition (one source tree) were highest for the coalescent bipartition (Table 1).
235 The MSWB showed 76,792 sites supported the supermatrix resolution, 90,659 sites
236 supported the coalescent resolution, and 19,575 sites supported the third bipartition (Fig.
237 3C). When examining only the sites where the MSWB was 2 log-likelihood units better
238 than the second best, 144 sites supported the supermatrix resolution, no sites supported
239 the coalescent resolution, and 16 sites supported the third bipartition.

240 For the carnivory dataset, we found 168 unique tree topologies gene, supermatrix,
241 and coalescent trees. The 168 tree topologies contained 41 conflicting bipartitions for the
242 relationship in question with 3 dominant bipartitions. The MGWB analyses found 500
243 genes supported the supermatrix bipartition, 498 genes supported the coalescent
244 bipartition, and 239 genes supported 16 additional bipartitions (Figs. 2D, 3E). The AIC

245 scores for the supermatrix (44 source trees), the coalescent (56 source trees), and for the
246 third bipartition (24 source trees) were highest for the supermatrix bipartition with genes
247 and the coalescent bipartition with sites (Table 1). The MSWB showed that 272,137 sites
248 supported the supermatrix bipartition, 615,418 sites supported the coalescent bipartition,
249 and 1,230,028 sites supported 18 additional bipartitions (Fig. 3F). When examining only
250 sites where the MSWB was 2 log-likelihood units better than the second best, 270 sites
251 supported the supermatrix topology, no sites supported the coalescent topology, and 34
252 sites supported two other topologies.
253
254 *Outlier gene examination*
255 For the carnivory dataset, we explored the possibility that the strongly conflicting genes
256 cluster575 and cluster3300 reflected methodological error in the assembly pipeline, as is
257 the case for the genes identified by Brown and Thomson (2017). However, both the
258 alignment and inferred phylogram for each gene revealed no obvious problems or
259 potential sources of systematic error (sparse alignment, abnormally long branch lengths,
260 etc.). We also explored whether compositional heterogeneity could explain the strongly
261 conflicting results (i.e., that the relationships were not truly conflicting, but instead
262 incorrectly modeled). However, both RY-coding in RAxML and explicit modeling of
263 multiple equilibrium frequencies (2, 3, or 4 composition regimes) across the tree in p4
264 v1.0 (Foster 2004) failed to overturn the inferred relationships. We further explored the
265 possibility of misidentified orthology. By examining the homolog tree produced from
266 amino acid data, we identified the ortholog from the nucleotide data to be complete (i.e.,
267 an ortholog within the homolog amino acid tree). We found that with the slower amino

acid data the sequences in the nucleotide cluster575 were inferred as a single monophyletic ortholog within a duplicated homolog (Supplementary Fig. 2). The discrepancies that appeared between the amino acid dataset and the CDS dataset were found to be either different in-paralogs/splice sites maintained during the dataset cleaning procedure or short sequences that were not identified as homologs in the CDS dataset (Supplementary Table 2 and Supplementary Fig. 2).

DISCUSSION

Biological processes including substitution saturation, hybridization, horizontal gene transfer, and incomplete lineage sorting can contribute to conflicting signal and may explain both conflict and lack of support widely found in phylogenomic datasets (Salichos et al. 2014; Smith et al. 2015; Kobert et al. 2016). In addition to these processes, limited taxonomic coverage for each gene is also a factor. While the carnivory dataset had extensive data overlap, the vertebrate dataset only had five gene regions that contained sequence data for every species (Supplementary Table 1). Another challenge for phylogenomic datasets is that high support values can mask significant underlying conflict (Ryan et al. 2013; Salichos et al. 2014; Smith et al. 2015; Kobert et al. 2016; Pease et al. 2017). Both datasets examined here recovered high support for two different topologies depending on the inclusion or exclusion of two genes with disproportionate influence on the likelihood (Brown and Thomson 2017; Walker et al. 2017). In the case of the carnivory dataset, the inferred topology changes with the inclusion or exclusion of just 0.0016% of the genes.

290 Outlying genes may be the result of biological processes or methodological errors.
291 The outlier genes in the vertebrate dataset were found to be the result of errors in
292 orthology detection and not biological processes (Brown and Thomson 2017). Here, we
293 explored the potential sources of error for the carnivorous outlier genes. While the
294 genomic resources are not available to fully examine the carnivorous outlier genes (e.g.,
295 we do not yet have synteny or information on gene loss), our analyses did not detect any
296 obvious problems with alignment, compositional heterogeneity, or homology. We found
297 one gene, cluster 575, to be an ortholog of a gene that experienced a duplication prior to
298 the divergence of both ingroup and outgroup taxa (Supplementary Fig. 2). While we
299 cannot rule out every possible source of error, we also cannot identify a source of
300 methodological error, suggesting the possibility that the conflicting topology is the result
301 of real (albeit unknown) biological processes.
302 The two-topology examination explored by most authors (Castoe et al. 2009;
303 Smith et al. 2011; Brown and Thomson 2017; Shen et al. 2017), whether performed with
304 Bayes factors or maximum likelihood, can be very informative and has been shown to
305 help identify these outlier genes. However, when questions involve a specific relationship,
306 the two-topology examination may not be optimal. The two-topology analyses require
307 fully bifurcating trees, while the research question may surround the resolution of one
308 relationship. Conflict among gene trees is a reality and expected from processes such as
309 incomplete lineage sorting. For instance, Jarvis et al. (2014) reported that no gene trees
310 from a genomic data set of 48 species of birds matched the inferred species tree. The
311 results of a two-topology analysis may be driven by the resolution of a part of the

312 phylogeny other than the area of interest, as two topology analyses condition on fully
313 bifurcating trees that necessarily resolve conflict in the entire tree.
314 To overcome these limitations, instead of full topologies, we examined
315 bipartitions across a set of candidate topologies, as defined by the set of gene trees and
316 the two tree hypotheses in question. By examining bipartitions, we accommodate for
317 uncertainty across the rest of the tree, regardless of the process generating that
318 uncertainty. For both datasets, both the MGWB and MSWB analyses demonstrate
319 increased support for the coalescent resolution as well as support for alternative
320 bipartitions. The vertebrate dataset contained three alternative bipartitions for the
321 relationship of interest while the carnivory dataset contained 41 different bipartitions
322 representing 168 topologies. The AIC analysis of the vertebrate dataset MGWBs and
323 MSWBs also suggested better fit of the coalescent bipartition than the supermatrix
324 bipartition. However, while the AIC analysis of MSWBs for the carnivorous dataset
325 showed better fit for the coalescent resolution, the AIC analysis of the MGWBs
326 supported the supermatrix resolution. This may be due to a higher penalty for parameters
327 for the coalescent bipartition as this included 56 gene trees as compared to the
328 supermatrix bipartition with 44 gene trees, but could also be the result of some within
329 gene processes driving conflicting signal.
330 The MSWB analyses comparing bipartitions that were 2 log likelihood better
331 than the second best provide additional information on the underlying data. In these
332 analyses, and for both datasets, the supermatrix resolution was supported. This is in
333 contrast to the results when considering all sites, not just those with 2 log likelihood
334 better likelihoods, and the AIC results for MSWBs that both support the coalescent

resolution. This suggests that in addition to outlier genes, there may be outlier sites that drive the resolution of these topologies in supermatrix analyses.

The results presented here contribute to a growing body of literature that address the question of how phylogenomic analyses should proceed in the presence of highly influential outlier genes, conflicting topologies, and ever expanding datasets (Wickett et al. 2014; Pease et al. 2016; Brown and Thomson 2017; Shen et al. 2017; Yang et al. 2017). For example, some authors have noted, and it is the case here, that supermatrix analyses may be more susceptible to the problem of strong outliers (Shen et al. 2017; Walker et al. 2017). In these cases, often the resolutions inferred using a coalescent method were favored. When the dominant process generating gene tree conflict is incomplete lineage sorting, coalescent methods should perform better (i.e., when gene tree diversity if modeled correctly). Some coalescent methods that weigh all gene tree equally (e.g., Mirarab and Warnow 2015), may overcome the problem of outlier genes even if incomplete lineage sorting is not the dominant source of conflict simply by eliminating the disproportionate influence of one or two outlying genes. Here, we demonstrate with two empirical examples that the coalescent resolution had higher support when examining bipartitions without using an explicit coalescent method. While we continue to learn about the patterns and processes that generate conflict and signal within phylogenetic datasets, new approaches that explore that data in different ways will continue to provide new insights. Further investigation into the behavior of each of these methods under different evolutionary processes is warranted to better understand their performance to the expected heterogeneity within transcriptomes or genomes.

FUNDING

JFW was 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 Nathanael Walker-Hale, Javier Igea, 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.

Brown J.M., Thomson R.C. 2017. Bayes Factors Unmask Highly Variable Information Content , Bias , and Extreme Influence in Phylogenomic Analyses. Syst. Biol. 66:517–530.

Brown J.W., Walker J.F., Smith S.A. 2017. Phyx: phylogenetic tools for unix. Bioinformatics. 33:1886–1888.

Castoe T.A., de Koning A.P.J., Kim H.-M., Gu W., Noonan B.P., Naylor G., Jiang Z.J., Parkinson C.L., Pollock D.D. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. 106:8986–8991.

Chiari Y., Cahais V., Galtier N., Delsuc F. 2012. Phylogenomic analyses support the

381 position of turtles as the sister group of birds and crocodiles (Archosauria). BMC
382 Biol. 10:65.
383 Foster P.G. 2004. Modeling compositional heterogeneity. Syst Biol. 53:485–495.
384 Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C.,
385 Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li
386 H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S.,
387 Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J.,
388 Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E.,
389 Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P.,
390 Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C. V., Lovell P. V., Wirthlin
391 M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M. V., Alfaro-
392 Nunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield
393 P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B.,
394 Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang
395 Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L.,
396 Barker F.K., Jonsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D.,
397 Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J.,
398 Burt D., Ellegren H., Alstrom P., Edwards S. V., Stamatakis A., Mindell D.P.,
399 Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-
400 genome analyses resolve early branches in the tree of life of modern birds. Science
401 (80-.). 346:1320–1331.
402 Kobert K., Salichos L., Rokas A., Stamatakis A. 2016. Computing the internode certainty
403 and related measures from partial gene trees. Mol. Biol. Evol. Advance Ac:1–17.

404 Mirarab S., Warnow T. 2015. ASTRAL-II: Coalescent-based species tree estimation with
405 many hundreds of taxa and thousands of genes. Bioinformatics. 31:i44–i52.
406 Morgan C.C., Foster P.G., Webb A.E., Pisani D., McInerney J.O., O'Connell M.J. 2013.
407 Heterogeneous models place the root of the placental mammal phylogeny. Mol. Biol.
408 Evol. 30:2145–56.
409 Pease J.B., Brown J.W., Walker J.F., Hinchliff C.E., Smith S.A. 2017. Quartet Sampling
410 distinguishes lack of support from conflicting support in the plant tree of life.
411 BioRxiv.
412 Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics Reveals Three
413 Sources of Adaptive Variation during a Rapid Radiation. PLoS Biol. 14:1–24.
414 Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Moriarty Lemmon E.,
415 Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted
416 next-generation DNA sequencing. Nature. 526:569–573.
417 Romiguier J., Ranwez V., Delsuc F., Galtier N., Douzery E.J.P. 2013. Less is more in
418 mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the
419 root of placental mammals. Mol. Biol. Evol. 30:2134–44.
420 Ryan J.F., Pang K., Schnitzler C.E., Nguyen A.D., Moreland R.T., Simmons D.K., Koch
421 B.J., Francis W.R., Havlak P., Smith S.A., Putnam N.H., Haddock S.H., Dunn C.W.,
422 Wolfsberg T.G., Mullikin J.C., Martindale M.Q., Baxevanis A.D. 2013. The genome
423 of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution.
424 Science (80-.). 342:1242592.
425 Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for
426 quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.

427 Shen X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic
428 studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:1–10.
429 Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets
430 reveals conflict, concordance, and gene duplications with examples from animals
431 and plants. BMC Evol. Biol. 15:150.
432 Smith S.A., Wilson N.G., Goetz F.E., Feehery C., Andrade S.C.S., Rouse G.W., Giribet
433 G., Dunn C.W. 2011. Resolving the evolutionary relationships of molluscs with
434 phylogenomic tools. Nature. 480:364–367.
435 Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-
436 analysis of large phylogenies. Bioinformatics. 30:1312–1313.
437 Walker J.F., Yang Y., Moore M.J., Mikenas J., Timoneda A., Brockington S.F., Smith
438 S.A. 2017. Widespread paleopolyploidy , gene tree conflict , and recalcitrant
439 relationships among the. Am. J. Bot. 104:858–867.
440 Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N.,
441 Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R.,
442 Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis
443 P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson
444 D.W., Surek B., Villarreal J.C., Roure B., Philippe H., DePamphilis C.W., Chen T.,
445 Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y.,
446 Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S., Leebens-Mack J. 2014.
447 Phylotranscriptomic analysis of the origin and early diversification of land plants.
448 Proc. Natl. Acad. Sci. 111:E4859–E4868.
449 Xi Z., Liu L., Rest J.S., Davis C.C. 2014. Coalescent versus Concatenation Methods and

450 the Placement of Amborella as Sister to Water Lilies. Syst. Biol. 63:919–932.

451 Yang Y., Moore M.J., Brockington S.F., Mikenas J., Olivieri J., Walker J.F., Smith S.A.

452 2017. Improved transcriptome sampling pinpoints 26 paleopolyploidy events in

453 Caryophyllales, including two paleo-allopolyploidy events. bioRxiv.

454 Zanis M.J., Soltis D.E., Soltis P.S., Mathews S., Donoghue M.J. 2002. The root of the

455 angiosperms revisited. Proc. Natl. Acad. Sci. U. S. A. 99:6848–53.

456

457

458 **Figure 1. Maximum likelihood trees inferred by RAxML for the Chiari et al. 2012**
459 **(vertebrate) and Walker et al. 2017 (carnivorous Caryophyllales) datasets.** Conflict
460 analysis for the vertebrate (A) and carnivory (B) datasets. The vertebrate analysis
461 includes the 93 genes that contained the outgroup (*Protopterus*), and the carnivory
462 analysis includes 1237 genes all of which had the outgroups (*Spinacia oleraceae* and
463 *Beta vulgaris*). Blue represents gene trees that are concordant with the relationship, grey
464 represents uninformative genes (SH-like < 80 or no taxon representation for the
465 bipartition), green represents the dominant alternate topology, and red represents all other
466 conflict. Numbers on edges represent concordance/conflict. Bold numbers at the nodes of
467 the vertebrate dataset correspond to bipartition numbers in Supplementary Table 1.
468

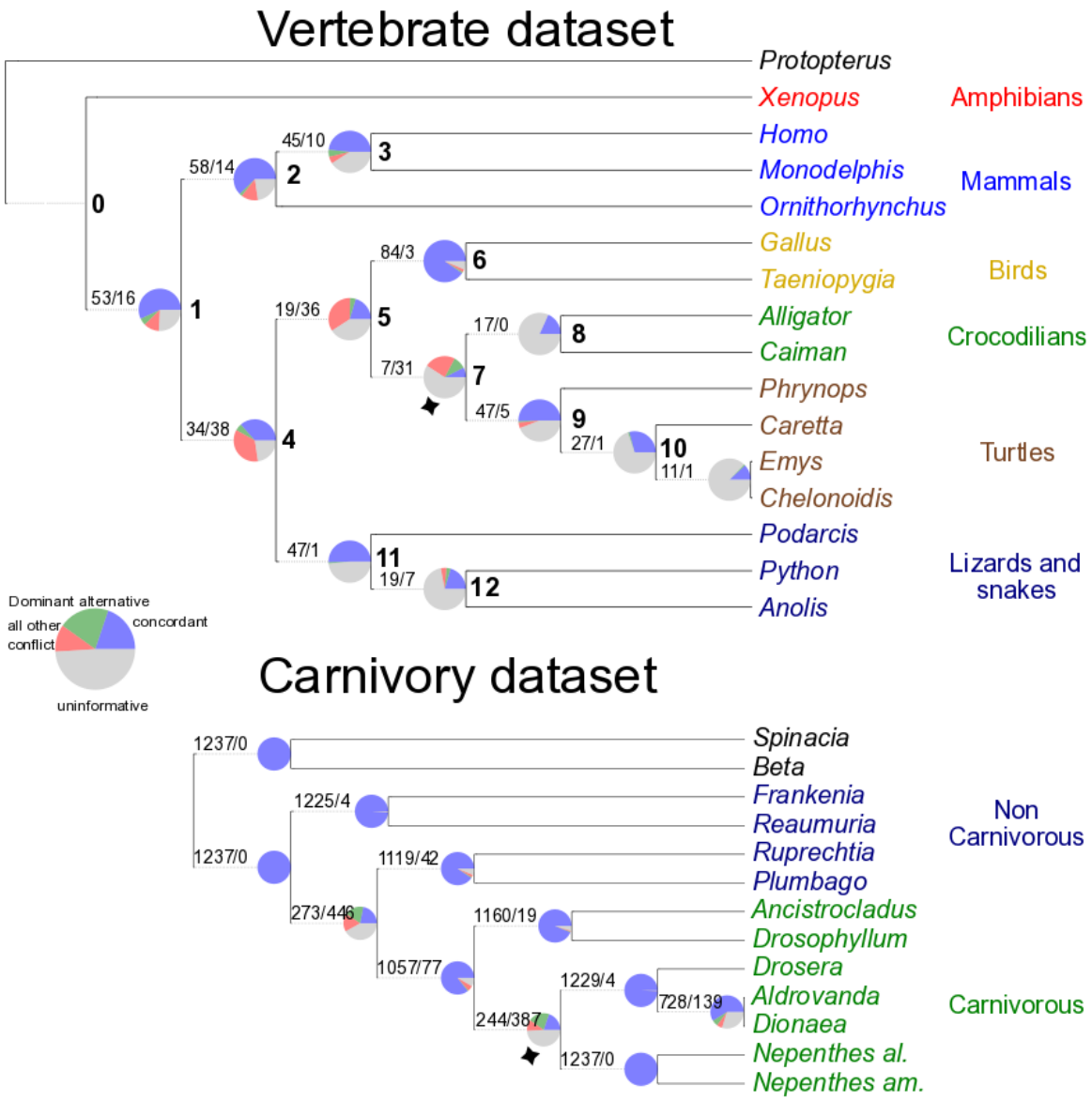


Figure 2. Comparison of two-topology gene-wise likelihood test and MGWB. A&C) Show the results of the two-topology test on the vertebrate and carnivory dataset, respectively, using the coalescent and supermatrix topologies as the comparison. B&D) Show the results of the MGWB test on the vertebrate and carnivory dataset, respectively, using the bipartitons found in the supermatrix topology, the coalescent topology, and unique gene trees with full species coverage.

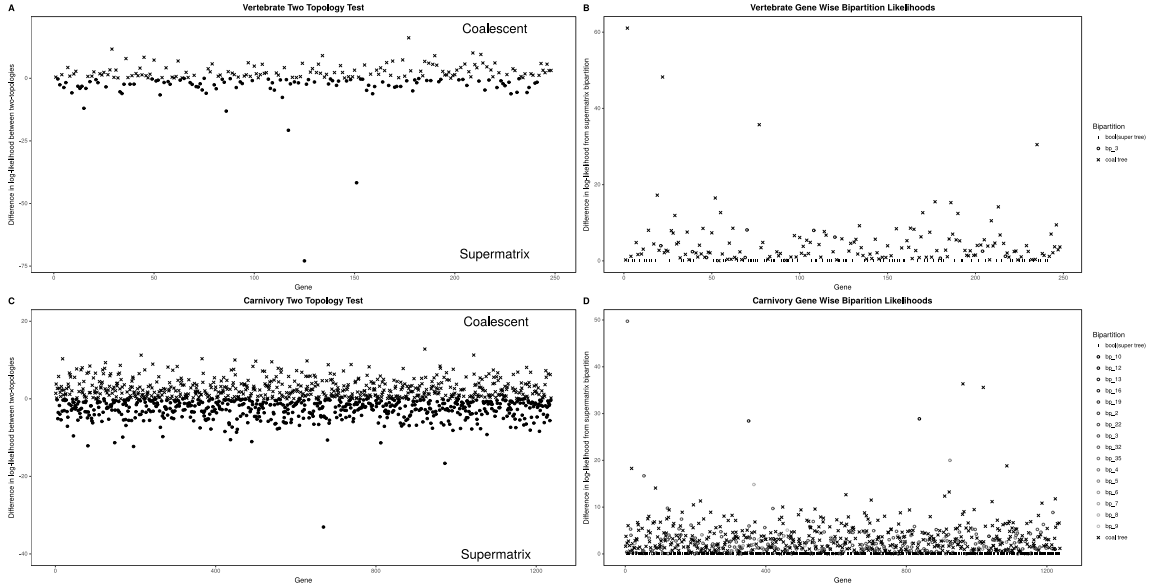


Figure 3. Bar plot representing the two-topology, MGWB, and MSWB methods for identifying support at a node. The letters S, C, and DA represent supermatrix, coalescent, and the dominant alternative bipartitions, respectively. Panels A-C show the results of the different methods applied to the vertebrate dataset with the most common dominant alternative shown in B&C having a clade consisting solely of mammals+crocodilians+turtles. Panels D-F show the results of the different methods applied to the carnivory dataset with the most common dominant alternative shown in E&F having a clade consisting of the two species in the genus *Nepenthes* sister to all other carnivorous Caryophyllales.

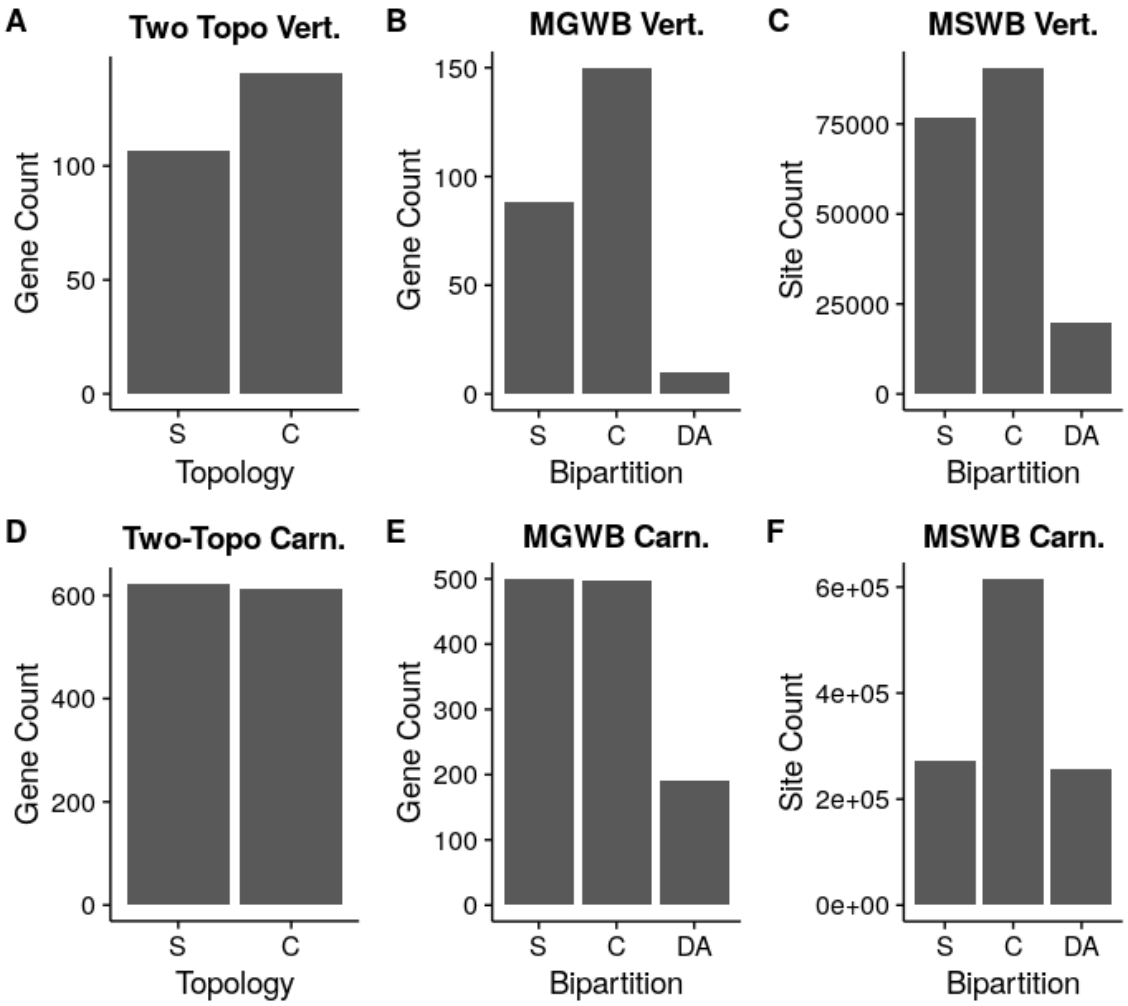


Table 1. Results of model testing the various topologies.

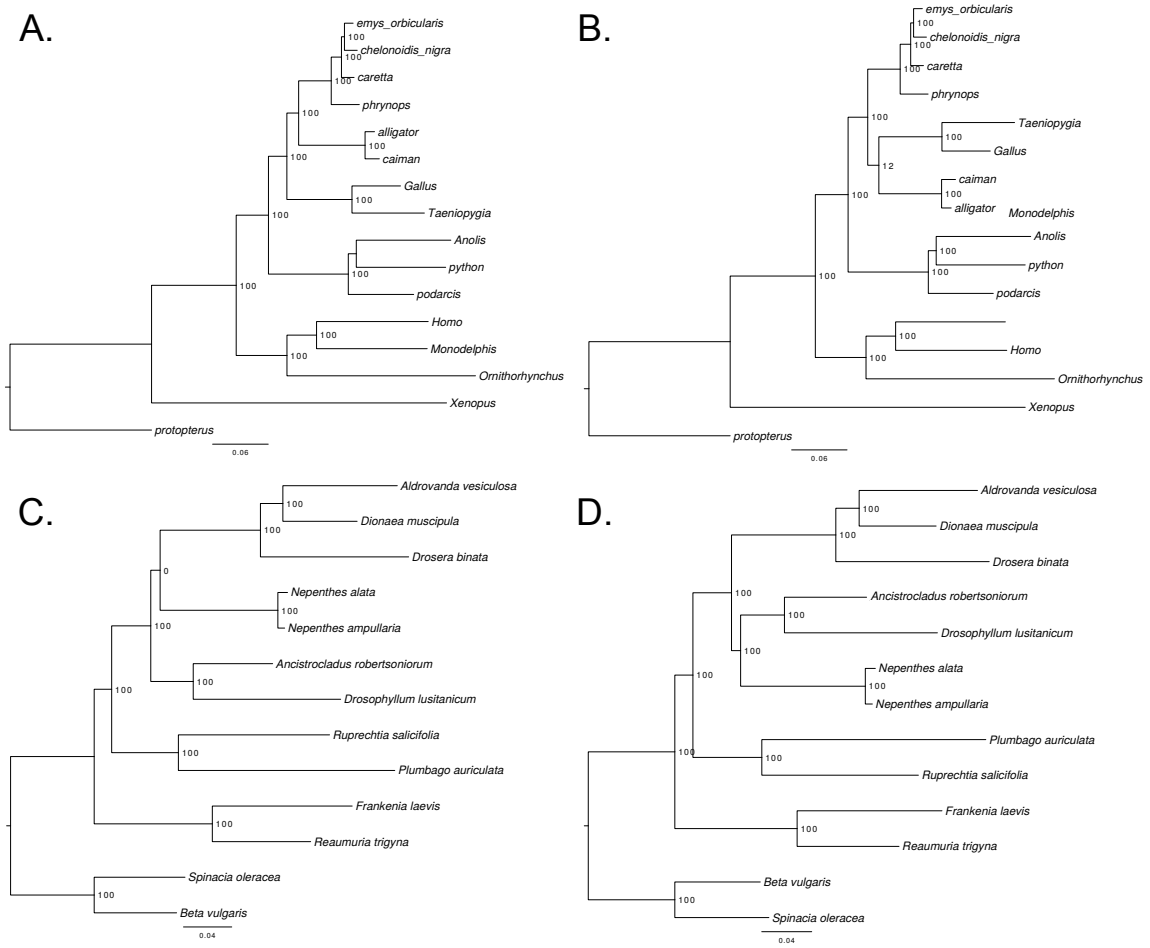
496

Comparison	Biparitition/Topology	AIC	ΔAIC	AIC_w
		Vertebrate Dataset		
Two Topology	Supermatrix	2094518	31	1.85e-07
	Coalescent	2094487	0.0	9.99e-01
MGWB	Supermatrix	2093604.12	622.46	0.0
	Coalescent	2092981.66	0.0	1.0
	Dominant Alternative	2098006.78	5025.12	0.0
MSWB	Supermatrix	2076738.66	2753.57	0.0
	Coalescent	2073985.09	0.0	1.0
	Dominant Alternative	2102131.54	28146.45	0.0
		Carnivory Dataset		
Two Topology	Supermatrix	26625000	0.0	1.0
	Coalescent	26625150	150	0.0
MGWB	Supermatrix	26611048.76	0.0	1.0
	Coalescent	26611339.93	291.17	0.0
	Dominant Alternative	26613649.05	2600.29	0.0
MSWB	Supermatrix	26032491.44	181865.6	0.0
	Coalescent	25850625.84	0.0	1.0
	Dominant Alternative	26101950.85	251325.01	0.0

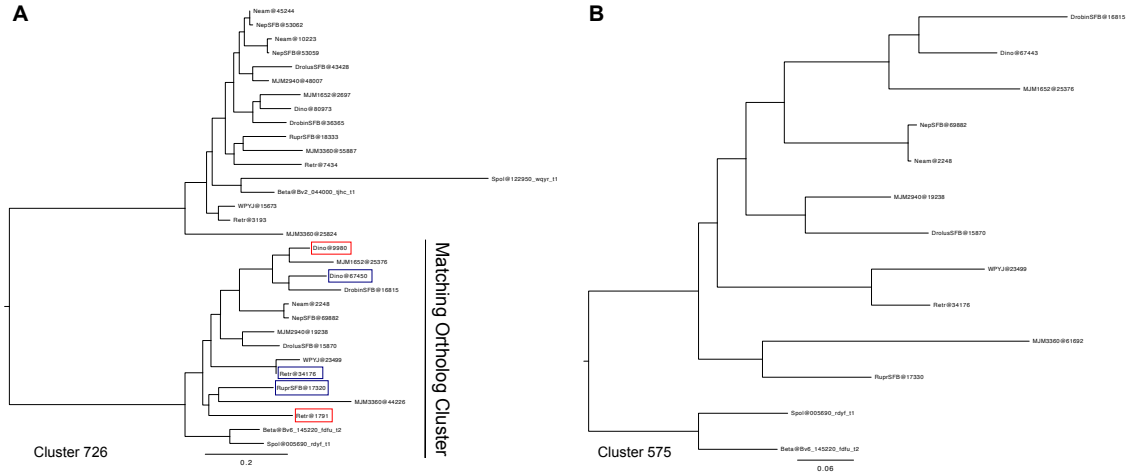
497

APPENDICES

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.



Supplementary Figure 2. Homolog tree for Amino Acid clustered (726) and CDS clustered (575) highly influential gene in the carnivorous Caryophyllales dataset. Different genes identified in the ortholog clusters are circled on cluster 726. Genes circled in red represent ones that are shorter and were not identified as orthologous in the CDS dataset and genes circled in blue represent alternate paralogs or introsplice sites used between the two clustering analyses.



519 **Supplementary Table 1.** Number of gene trees in which all the species for a given

520 bipartition are present. Bipartitions correspond to node labels on Fig. 1.

Bipartition number	Genes containing all species for the bipartition
0	5
1	5
2	246
3	248
4	5
5 (All turtle, crocodilians, and birds)	6
6	248
7	6
8	23
9	36
10	45
11	69
12	51
13	94
Bipartition of turtles sister to birds+crocodilians	36

521

522

523 **Supplementary Table 2. Sources of discrepancy between the orthologs detected in**

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

525 **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

526