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 supermatrix analyses. In this study, we examined plant and vertebrate datasets 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 a small number of fixed topologies, often two topologies, gene tree topologies generated under processes such as incomplete lineage sorting are unlikely to precisely match these topologies. We therefore examined edges across a set of trees and recover more support for the resolution favored by coalescent analyses. These results suggest that by expanding beyond fixed-topology comparisons, we can dramatically improve our understanding of the underlying signal in phylogenomic datasets by asking more targeted edge-based questions.

INTRODUCTION

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 within 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 "outlier genes" on a supermatrix analysis is capable of altering tree topology inference (Brown and Thomson 2017; Shen et al. 2017)(Brown and Thomson 2017; Shen et al. 2017)(Brown and Thomson 2017; Shen et al. 2017). Using a Bayes factor approach Brown and Thomson (2017) reanalyzed a series of published datasets 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

63 birds with 100% PP. The same topology was previously inferred using ML with
64 nucleotide data in the original study by Chiari et al. (2012), but was dismissed in favor of
65 a coalescent reconstruction that placed turtles sister to birds+crocodilians. When Brown
66 and Thomson (2017) removed the outlier genes, the reduced supermatrix infers the same
67 topology as the coalescent reconstruction with 100% PP. Another recently published
68 study compared gene-wise likelihoods across multiple topologies to examine contentious
69 relationships across the tree of life and found disproportionate influence of genes at all
70 contentious relationships examined (Shen et al. 2017).

71 Given the prevalence of outlier genes in phylogenomic datasets, and the continued
72 focus on contentious relationships in the tree of life, it is imperative that we develop
73 methods for analyzing conflict and selecting among alternative resolutions for recalcitrant
74 relationships. We build upon the discussions of Brown and Thomson (2017) and Shen et
75 al. (2017) by addressing whether these outlier genes violate models of evolution.
76 Furthermore, we present a method that expands on topology comparisons to instead
77 pursue edge-based questions. Typically, site-wise and gene-wise log-likelihood analyses
78 of phylogenomic datasets are performed in a pairwise manner on two or more fixed
79 alternate topologies (e.g., Castoe et al. 2009; Smith et al. 2011; Shen et al. 2017).
80 However, given widespread gene tree discordance (e.g., due to incomplete lineage
81 sorting), it may be more realistic to assume that many alternative topologies are
82 supported within larger genomic datasets (e.g., Smith et al. 2015; Pease et al. 2016;
83 Walker et al. 2017). Additionally, when the research question involves a single
84 relationship and not the entirety of the tree, it may be more appropriate to examine
85 targeted edges instead of resolved topologies (Lee and Hugall 2003). This allows for any

processes that may be causing conflict in the non-focal parts of the tree to be accommodated without influencing the relationships of interest. Here, we compare results from two-topology gene-wise log-likelihood analyses and a novel approach of gene-wise edge (MGWE) analysis (see Methods below). We examine 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. In both cases we find that the use of an edge based approach results in stronger support for the topology hypothesized to be correct by researchers in the original study.

METHODS

Data collection

We obtained the 248 genes that were codon-aligned and analyzed by Brown and Thomson (2017) from the 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 and the code to conduct the MGWE analysis may be found at https://github.com/jfwalker/SiteSpecificLogLikelihood.

109

110 *Species trees*

111 Brown and Thomson (2017) used Bayesian analyses to obtain the topologies from the
112 Chiari et al. (2012) data set. As our study focused on the use of maximum likelihood
113 (ML) for detecting overly influential genes, we ensured that ML phylogenetic
114 reconstruction would recapitulate the previous species tree results. To construct a
115 supermatrix tree for the vertebrate dataset, the 248 individual vertebrate genes used in
116 Brown and Thomson (2017) were concatenated using the Phyx program pxcat (Brown et
117 al. 2017). The species tree was inferred in RAxML v8.2.1 (Stamatakis 2014) using the
118 GTR+ Γ model of evolution, and edge support was assessed from 200 rapid bootstrap
119 replicates. Supermatrix trees for the vertebrate dataset were inferred both with all genes
120 present, and again with the previously identified two outlier genes (8916 and 11434)
121 removed (see below). The ML tree inferred from all the data from the carnivory dataset
122 was downloaded from (http://dx.doi.org/10.5061/dryad.33m48) while a novel ML tree
123 was inferred from a reduced supermatrix that excluded two highly informative genes
124 (cluster575 and cluster3300; see below).

125

126 *Gene tree construction and analysis of conflict*

127 Individual gene trees for both datasets were inferred using ML with the GTR+ Γ model of
128 evolution as implemented in RAxML. A SH-like test (Anisimova et al. 2011), as
129 implemented in RAxML, was performed to assess gene tree edge support. As this test
130 examines alternative topologies by nearest-neighbor interchange (NNI), it is possible that
131 during the test a topology with a higher likelihood is found (i.e., an 'NNI-optimal'

132 topology). When a better topology was found during the test performed for this study,
133 that topology was used in downstream analyses. We used the pxrr program in the Phyx
134 package (Brown et al. 2017) to root all gene trees on the outgroup (*Protopterus* for the
135 vertebrate dataset, and *Beta vulgaris* and *Spinacia oleraceae* for the carnivory dataset)
136 and we excluded gene trees where an outgroup was not present. We mapped conflict onto
137 the supermatrix tree using phyparts (Smith et al. 2015) with SH-like support of < 80
138 treated as uninformative. We chose 80 as a support cutoff due to the traditional cutoff of
139 (95) being shown as overly conservative with this test (Guindon et al. 2010). Gene tree
140 conflict was visualized using the script phypartspiecharts.py (available from
141 https://github.com/mossmatters/MJPythonNotebooks). We conducted more detailed
142 conflict analyses used for edge comparisons discussed below using pxbp as part of the
143 Phyx package (Brown et al. 2017).
144
145 *Calculating two-topology gene-wise log-likelihoods*
146 The alternate topologies (supermatrix and coalescent) and data matrices for the vertebrate
147 and carnivory datasets were obtained from the original studies, Chiari et al. (2012) and
148 Walker et al. (2017), respectively. We calculated site-wise log-likelihood scores for the
149 two topologies in RAxML using the GTR+ Γ model of evolution, with the data
150 partitioned by gene. The differences in site-wise log-likelihoods between the candidate
151 topologies were then calculated using scripts available from
152 https://bitbucket.org/jfwalker/maximizelikelihoods and
153 https://github.com/jfwalker/SiteSpecificLogLikelihood.
154

Maximum gene-wise edge calculations

In addition to pairwise topological comparisons, we also examined the maximum gene-wise edges (MGWE) (Fig 1.). For a single gene and a single focal edge, the MGWE is the likelihood of a gene tree with the highest likelihood that also displays the edge of interest. When calculating the MGWE for a focal edge across multiple genes, this approach does not require each gene to have the same topology, just that the likelihood comes from a tree that displays the edge of interest. This contrasts with a standard fixed topology comparison where the topology for each gene would be required to be the same (e.g., supermatrix vs. coalescent topology). Unlike the fixed topology approach the MGWE allows for genes to have conflicting relationships outside of the edge of interest. Here, we are interested in comparing the MGWE for sets of alternative and conflicting edges in order to determine if, by relaxing the requirement for each gene to share the topology, we gain insight into the signal for conflicting relationships. One could calculate the MGWE on any number of edges, and we consider the dominant alternative edges as identified in the literature.

While there are several ways that MGWEs could be calculated, we restricted the tree space under under consideration by circumscribing a set of empirically-supported topologies (TREESET) consisting of the supermatrix-inferred topology, coalescent inferred topology, and individual gene trees that contained all taxa. We then identified the conflicting trees and pooled trees based upon shared conflicting relationships for the edges of interest (EDGE). We then calculated the maximum likelihood for each gene and for each topology.

For the edges of interest, we calculated the MGWEs by retaining the likelihood for the tree with the highest likelihood that displayed the focal EDGE. This became the representative likelihood for that EDGE. We then summed the representative likelihoods together. That value, however, is not comparable between edges because a different number of trees may be compared (Theobald 2010). Therefore, we calculated AIC scores (-2ln(L) + 2k) for each EDGE. This, effectively, allowed for comparisons between more parameter rich models and parameter poor models. The parameters, k, were calculated based on the number of taxa, n, and the number of genes in the analysis, g. The branch length parameters equal $2\times n - 3$ and the GTR + Γ model of evolution = 6 (where base frequencies were empirical and not estimated). The supermatrix ML analyses that assumed a single set of branch lengths on one topology and model parameters to be unlinked across genes consisted of $2\times n - 3 + 6\times g$ parameters. For each EDGE, because branch lengths were calculated for each gene tree, the parameters consisted of the sum of the number of parameters used for each gene: g x (2 x n – 3 + 6). In addition to calculating AICs for the coalescent and supermatrix topologies with a single set of branch lengths across the gene set, we also calculated AICs allowing the branch lengths to vary across genes. This calculation results in the same number of parameters as the EDGE calculations. Here, we are focused on addressing conflicting signal between edges of interest and so the increase in the number of parameters is acceptable considering our examination of gene trees. However, future work could attempt to limit the expansion of the number of parameters for each EDGE by sharing branch length estimates or model parameters across genes.

200 *Testing for paralogy in carnivory dataset*

201 The homolog trees created from amino acid data in the study by Walker et al. (2017)
202 were downloaded from Dryad (http://datadryad.org/resource/doi:10.5061/dryad.vn730).
203 We matched the sequences from the outlier genes to their corresponding sequence in the
204 amino acid homolog trees. This allowed us to examine whether a nucleotide cluster
205 contained homology errors that may be exposed by the slower evolving amino acid
206 dataset.

207

208 RESULTS

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

210 Our ML analysis of the vertebrate dataset recovered the same supermatrix topology (Fig.
211 2) as found with ML by Chiari et al. (2012) and Bayesian inference by Brown and
212 Thomson (2017). The difference in log-likelihood between the supermatrix and
213 coalescent topologies for the vertebrate dataset was 4.01. Ninety-three of 248 gene trees
214 could be rooted on the outgroup *Protopterus* and only five of these had all taxa
215 represented (Supplementary Table 1). We found low support for relationships within
216 gene trees (SH <80) and significant gene tree conflict (Fig. 2). Of the gene trees with high
217 support (SH >80), seven resolved turtles+crocodilians as sister to birds (hereafter referred
218 to as the vertebrate supermatrix topology) and nine resolved crocodilians+birds sister to
219 turtles (hereafter referred to as the vertebrate coalescent topology).

220 The two-topology gene-wise log-likelihood comparison showed that 105 genes
221 had a higher likelihood score for the vertebrate supermatrix topology while 143 supported
222 the vertebrate coalescent topology (Figs. 3A, 4A). Two genes (ENSGALG00000008916

223 and ENSGALG00000011434, referred to here as 8916 and 11434, respectively),
224 appeared as outliers, exhibiting a disproportionate influence on the overall likelihood of
225 the supermatrix (Fig. 3A). The outlier genes identified with maximum likelihood
226 analyses matched those previously identified as outliers using Bayes factors (Brown and
227 Thomson 2017). These two genes both supported the vertebrate supermatrix topology
228 with log-likelihood scores of 79.55 and 46.01 greater than the alternative coalescent tree
229 topology, respectively. The difference in log-likelihood between the two topologies of the
230 non-outlier genes ranged from |0.006| to |19.891| with an average of 3.31 for all genes in
231 the analysis. The removal of the vertebrate genes 8916 and 11434, as shown by Brown
232 and Thomson (2017), recovered the coalescent topology, albeit with low bootstrap
233 support (BS = 12; Supplementary Fig. 1).
234 Previous work on the carnivory dataset demonstrated that the placement of the
235 *Ancistrocladus+Drosophyllum* clade (Fig. 2) contained significant conflict and is
236 strongly influenced by species sampling (Walker et al. 2017). The log-likelihood
237 difference between the supermatrix and coalescent topologies was 74.94 in favor of the
238 former. The two-topology log-likelihood comparison between the dominant topologies on
239 the carnivory dataset (Fig. 3B) showed that 623 genes supported
240 *Ancistrocladus+Drosophyllum* sister to all other carnivorous plants (hereafter referred to
241 as carnivory supermatrix topology) while 614 genes supported
242 *Ancistrocladus+Drosophyllum* sister to *Nepenthes alata+Nepenthes ampullaria*
243 (hereafter referred to as carnivory coalescent topology; Figs. 3A & 4D). Two genes
244 (cluster575 and cluster3300) contributed disproportionately to the overall likelihood.
245 Individually these two genes have a difference in log-likelihood scores between the two

246 topologies of 33.06 and 16.63, respectively, and support the carnivory supermatrix
247 topology. When we reanalyzed the supermatrix with cluster575 and cluster3300 removed,
248 the carnivory coalescent topology was recovered, with 100% BS support (Supplementary
249 Fig. 1). The difference between the two topologies in log-likelihood of the non-outlier
250 genes ranged from |0.001| to |12.82| with an average of 2.82 for all genes in the analysis.
251

252 *Edge based analysis changes supported topology*

253 We compared MGWE and two topology gene-wise likelihoods involving the contentious
254 bird, crocodilian, and turtle relationships in the vertebrate dataset (Fig. 4B). We found
255 seven unique topologies with the necessary species coverage to conduct the analyses: five
256 gene tree topologies from Chiari et al. (2012) and the two dominant species tree
257 topologies. The set of seven trees included three major conflicting edges for the
258 relationship in question: the two resolutions found in the supermatrix and coalescent trees,
259 and birds sister to crocodilian+mammals+turtles. 91 genes supported the vertebrate
260 supermatrix edge, 144 genes supported the vertebrate coalescent edge, and 13 genes
261 supported the third conflicting edge (Fig. 4B). When comparing the supermatrix analysis
262 with a single set of branch lengths, to that where it was treated as a sum of gene tree
263 likelihoods, we found a superior AIC score for the sum of gene tree likelihoods (Table 1).
264 The MGWE AIC scores for the summed likelihoods of the supermatrix (three source
265 trees), the coalescent (three source trees), and the third conflicting edge (one source tree)
266 were highest for the coalescent edge and out of all tested models the coalescent edge was
267 inferred to be the best (Table 1).

268 For the carnivory dataset, we found 168 unique tree topologies to include in the
269 tree set. The 168 tree topologies contained 41 conflicting edges for the relationship in
270 question with 3 dominant edges. The MGWE analyses found 499 genes supported the
271 supermatrix edge, 466 genes supported the coalescent edge, and 272 genes supported 15
272 additional edges (Figs. 2D, 3E). When we further compared the MGWE AIC scores for
273 the supermatrix (44 source trees), the coalescent (56 source trees), and for the third edge
274 (24 source trees) we found the coalescent edge to have the best AIC score out of all tested
275 models (Table 1).
276
277 *Outlier gene examination*
278 For the carnivory dataset, we explored the possibility that the strongly conflicting genes
279 cluster575 and cluster3300 reflected methodological error in the assembly pipeline, as is
280 the case for the genes identified by Brown and Thomson (2017) for the vertebrate dataset.
281 However, both the alignment and inferred phylogram for each gene revealed no obvious
282 problems or potential sources of systematic error (sparse alignment, abnormally long
283 branch lengths, etc.). We also explored whether compositional heterogeneity could
284 explain the strongly conflicting results (i.e., that the relationships were not truly
285 conflicting, but instead incorrectly modeled). However, both RY-coding in RAxML and
286 explicit modeling of multiple equilibrium frequencies (2, 3, or 4 composition regimes)
287 across the tree in p4 v1.0 (Foster 2004) failed to overturn the inferred relationships. We
288 further explored the possibility of misidentified orthology. By examining the homolog
289 tree produced from amino acid data, we identified the ortholog from the nucleotide data
290 to be complete (i.e., an ortholog within the homolog amino acid tree). We found that with

291 the slower amino acid data the sequences in the nucleotide cluster575 were inferred as a
292 single monophyletic ortholog within a duplicated homolog (Supplementary Fig. 2). The
293 discrepancies that appeared between the amino acid dataset and the CDS dataset were
294 found to be either different in-paralogs/splice sites maintained during the dataset cleaning
295 procedure or short sequences that were not identified as homologs in the coding DNA
296 sequence (CDS) dataset (Supplementary Table 2 and Supplementary Fig. 2).
297

298 ## DISCUSSION

299 Biological processes including substitution saturation, hybridization, horizontal gene
300 transfer, and incomplete lineage sorting can contribute to conflicting signal and may
301 explain both conflict and lack of support widely found in phylogenomic datasets
302 (Salichos et al. 2014; Smith et al. 2015; Kobert et al. 2016). In addition to these
303 biological processes, other data set assembly issues such as limited taxonomic coverage
304 for each gene may also contribute to conflict and low support in these data sets. For
305 example, while the carnivory dataset had extensive data overlap, the vertebrate dataset
306 only had five gene regions that contained sequence data for every species (Supplementary
307 Table 1). To further complicate the challenges facing phylogenomic analyses, high
308 support values, especially from concatenated runs, can mask significant underlying
309 conflict (Lee and Hugall, 2003; Ryan et al. 2013; Salichos et al. 2014; Smith et al. 2015;
310 Kobert et al. 2016; Pease et al. 2017). Both datasets examined here recovered high
311 support for two different topologies depending on the inclusion or exclusion of two genes
312 with disproportionate influence on the likelihood (Brown and Thomson 2017; Walker et

313 al. 2017). In the case of the carnivory dataset, the inferred topology changes with the
314 inclusion or exclusion of just 0.0016% of the genes.

315 To address these challenges, several approaches have been outlined in the
316 literature. Recently, the discovery of outlier genes has resulted in the necessity to closely
317 examine gene tree topologies and likelihoods (Brown and Thomson 2017; Shen et al.
318 2017). Outlier genes may be the result of biological processes or methodological errors,
319 and due to their high influence of species tree inference should be thoroughly examined.
320 Previously, the outlier genes in a vertebrate dataset were found to be the result of errors
321 in orthology detection and not biological processes (Brown and Thomson 2017). We
322 explored, in this study, the potential sources of error for the outlier genes in a dataset of
323 carnivorous plants. While the genomic resources are not available to fully examine the
324 carnivorous outlier genes (e.g., we do not yet have synteny or information on gene loss),
325 our analyses did not detect any obvious problems with alignment, compositional
326 heterogeneity, or homology. We found one gene, cluster575, to be an ortholog of a gene
327 that experienced a duplication event prior to the divergence of both ingroup and outgroup
328 taxa (Supplementary Fig. 3). While we cannot rule out every possible source of error, we
329 also cannot identify a source of methodological error, suggesting the possibility that the
330 conflicting topology is the result of real (albeit unknown) biological processes.

331 Fixed topological and pairwise examinations explored by most authors (Castoe et
332 al. 2009; Smith et al. 2011; Shen et al. 2017), have been very informative for the
333 identification of not only outlier genes, but also for phylogenetic signal for and against
334 conflicting phylogenetic relationships. However, for many reasons, these fixed
335 topological examinations, where a single topology is assumed to underlie all genes, may

336 not be optimal. Conflict among gene trees is common and expected from processes such
337 as incomplete lineage sorting, hybridization, and other processes. For instance, Jarvis et
338 al. (2014) reported that no gene trees from a genomic data set of 48 species of birds
339 matched the inferred species tree. Furthermore, such a result becomes increasingly likely
340 as sampling breadth (both taxa within a clade as well as the age of the clade itself)
341 increases. The results of a fixed-topology analysis may be driven by the resolution of a
342 part of the phylogeny other than the area of interest, as fixed-topology analyses condition
343 on fully bifurcating trees that necessarily resolve conflict in the entire tree.

344 To overcome these limitations, instead of fixed singular topologies, we examined
345 edges across a set of empirically supported candidate topologies, as defined by the set of
346 inferred gene trees and the two tree hypotheses in question. By examining edges, we
347 accommodate for uncertainty across the rest of the tree, regardless of the process
348 generating that uncertainty. We examined this with both a vertebrate dataset and
349 carnivorous plants dataset discussed above. The vertebrate dataset contained three
350 alternative edges for the relationship of interest while the carnivory dataset contained 41
351 different edges representing 168 topologies. The MGWE analysis and AIC scores of both
352 the vertebrate dataset and the carnivory dataset both suggested a better fit of the
353 coalescent edge than the supermatrix edge (Table 1). Also, in both cases, we found that
354 the AIC score supported the higher parameterized model, as opposed to a single fixed
355 topology and branch lengths. While we do not suggest that this is the best fit model and
356 only the best of the ones analyzed here, this indicates that future studies may benefit from
357 allowing more heterogeneity than is typically involved in a concatenation analysis. This
358 will require careful examination of some of the complexity involved in these large

359 phylogenomic analyses. For example, there is the issue of how missing data is handled in
360 these calculations (e.g.,Stamatakis and Alachiotis 2010). Furthermore, the models
361 explored could potentially have significantly reduced parameters by sharing topologies
362 and branch lengths across some gene regions, including potentially scaling branch lengths
363 proportionally (e.g., as is possible with the -spp option in the program iqtree).
364 Nevertheless, the exploratory analyses presented here provide additional evidence that a
365 simple concatenation approach with these large datasets masks important heterogeneity
366 that can be analyzed further to help inform phylogenetic resolution.

367 The results presented here contribute to a growing body of literature that address
368 the question of how phylogenomic analyses should proceed in the presence of highly
369 influential outlier genes, conflicting topologies, and ever expanding datasets (Wickett et
370 al. 2014; Pease et al. 2016; Brown and Thomson 2017; Shen et al. 2017; Yang et al.
371 2017). For example, some authors have noted, and it is the case here, that supermatrix
372 analyses may be more susceptible to the problem of strong outliers (Shen et al. 2017;
373 Walker et al. 2017). In these studies, the resolutions inferred using a coalescent method
374 were generally favored. When the dominant process generating gene tree conflict is
375 incomplete lineage sorting, coalescent methods should perform better (i.e., when gene
376 tree diversity is modeled correctly). Some coalescent methods that weigh all gene tree
377 equally (e.g., Mirarab and Warnow 2015), may overcome the problem of outlier genes
378 even if incomplete lineage sorting is not the dominant source of conflict simply by
379 eliminating the disproportionate influence of one or two outlying genes. Here, we
380 demonstrate with two empirical examples that the coalescent resolution had higher
381 support when examining edges without using an explicit coalescent method.

While we continue to uncover the patterns and processes that generate conflicting signal within phylogenomic datasets, it is imperative that we explore new methods that accommodate conflict. Phylogenomic studies often focus sampling efforts around particularly recalcitrant nodes, and it is important we develop methods designed with the same purpose. Here we focus on conflicting edges and explore the MGWE method as a means of analyzing these conflicting edges while allowing for topological heterogeneity outside of the relationships of interest. This approach helps accommodate the biological realities of heterogeneity among lineages, conflicting signal both for in and outside the relationship of interest, and evolutionary processes that violate assumptions by supermatrix and coalescent models. This approach, however, is just a start and future research should examine how to better incorporate the underlying heterogeneity that has emerged from our large data sets over the last few years. We believe further investigation into edge based testing is warranted to better understand how we may incorporate the process based conflict of phylogenomics into our analyses.

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 Alexis Stamatakis for advice on RAxML and 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 position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10:65.

Foster P.G. 2004. Modeling compositional heterogeneity. Syst Biol. 53:485–495.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the

428 performance of PhyML 3.0. Syst. Biol. 59:307–321.
429 Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C.,
430 Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li
431 H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S.,
432 Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J.,
433 Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E.,
434 Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P.,
435 Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C. V., Lovell P. V., Wirthlin
436 M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M. V., Alfaro-
437 Nunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield
438 P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B.,
439 Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang
440 Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L.,
441 Barker F.K., Jonsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D.,
442 Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J.,
443 Burt D., Ellegren H., Alstrom P., Edwards S. V., Stamatakis A., Mindell D.P.,
444 Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-
445 genome analyses resolve early branches in the tree of life of modern birds. Science
446 (80-.). 346:1320–1331.
447 Kobert K., Salichos L., Rokas A., Stamatakis A. 2016. Computing the internode certainty
448 and related measures from partial gene trees. Mol. Biol. Evol. Advance Ac:1–17.
449 Lee M.S.Y., Hugall A.F. 2003. Partitioned Likelihood Support and the Evaluation of
450 Data Set Conflict. Syst. Biol. 52:15–22.

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

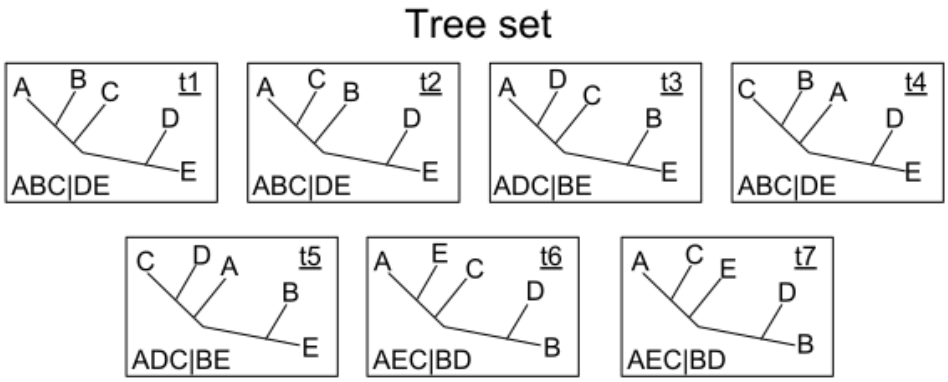
474 Shen X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic
475 studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:1–10.
476 Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets
477 reveals conflict, concordance, and gene duplications with examples from animals
478 and plants. BMC Evol. Biol. 15:150.
479 Smith S.A., Wilson N.G., Goetz F.E., Feehery C., Andrade S.C.S., Rouse G.W., Giribet
480 G., Dunn C.W. 2011. Resolving the evolutionary relationships of molluscs with
481 phylogenomic tools. Nature. 480:364–367.
482 Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-
483 analysis of large phylogenies. Bioinformatics. 30:1312–1313.
484 Stamatakis A., Alachiotis N. 2010. Time and memory efficient likelihood-based tree
485 searches on phylogenomic alignments with missing data. Bioinformatics. 26:132–
486 139.
487 Theobald D.L. 2010. A formal test of the theory of universal common ancestry. Nature.
488 465:219–222.
489 Walker J.F., Yang Y., Moore M.J., Mikenas J., Timoneda A., Brockington S.F., Smith
490 S.A. 2017. Widespread paleopolyploidy , gene tree conflict , and recalcitrant
491 relationships among the. Am. J. Bot. 104:858–867.
492 Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N.,
493 Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R.,
494 Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis
495 P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson
496 D.W., Surek B., Villarreal J.C., Roure B., Philippe H., DePamphilis C.W., Chen T.,

Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S., Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. 111:E4859–E4868.

Xi Z., Liu L., Rest J.S., Davis C.C. 2014. Coalescent versus Concatenation Methods and the Placement of Amborella as Sister to Water Lilies. Syst. Biol. 63:919–932.

Yang Y., Moore M.J., Brockington S.F., Mikenas J., Olivieri J., Walker J.F., Smith S.A. 2017. Improved transcriptome sampling pinpoints 26 paleopolyploidy events in Caryophyllales, including two paleo-allopolyploidy events. bioRxiv.

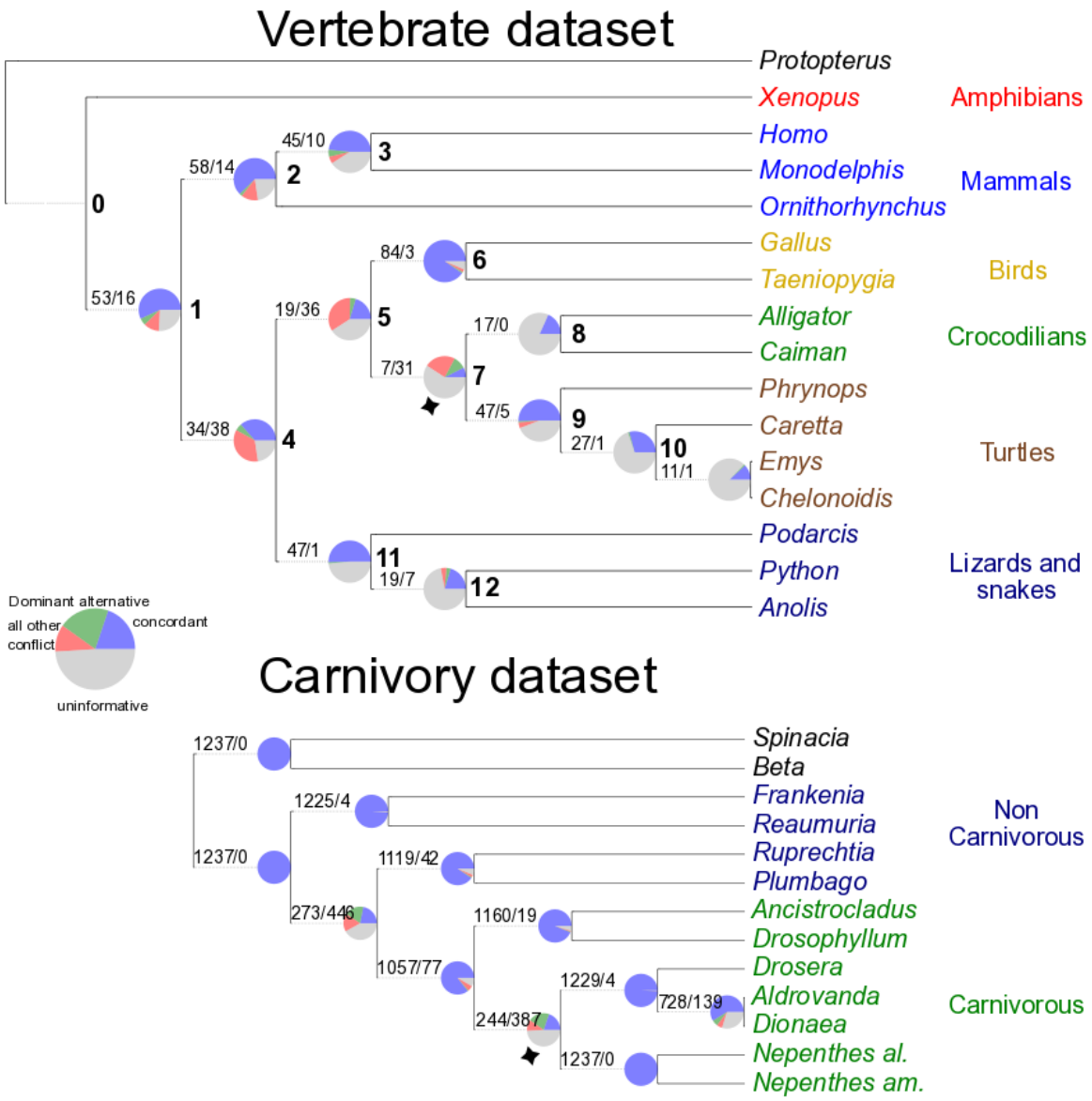
Zanis M.J., Soltis D.E., Soltis P.S., Mathews S., Donoghue M.J. 2002. The root of the angiosperms revisited. Proc. Natl. Acad. Sci. U. S. A. 99:6848–53.



MGWE

Edge	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
ABC | DE	-20.00 (t1)	-10.00 (t4)	-11.00 (t2)	-14.00 (t1)	-9.00 (t1)
ADC | BE	-90.00 (t3)	-5.00 (t5)	-4.00 (t5)	-50.00 (t3)	-20.00 (t5)
AEC | BD	-15.00 (t6)	-11.00 (t7)	-7.00 (t6)	-8.00 (t7)	-10.00 (t6)

Figure 1. Outline for the MGWE procedure. The inferred tree set is depicted at the top, with the tree number in the top right hand corner of each box, and the edge representing the relationship of interest in the bottom left hand corner. The MGWE shows the best likelihood for each edge at each gene, with the tree from which that likelihood was obtained in the box in parentheses next to the likelihood score.



524

525 **Figure 2. Maximum likelihood trees inferred by RAxML for the Chiari et al. 2012**

526 **(vertebrate) and Walker et al. 2017 (carnivorous Caryophyllales) datasets.** Conflict

527 analysis for the vertebrate (A) and carnivory (B) datasets. The vertebrate analysis

528 includes the 93 genes that contained the outgroup (*Protopterus*), and the carnivory

529 analysis includes 1237 genes all of which had the outgroups (*Spinacia oleraceae* and

530 *Beta vulgaris*). Blue represents gene trees that are concordant with the relationship, grey

531 represents uninformative genes (SH-like < 80 or no taxon representation for the edge),

green represents the dominant alternate topology, and red represents all other conflict. Numbers on edges represent concordance/conflict. Bold numbers at the nodes of the vertebrate dataset correspond to edge numbers in Supplementary Table 1.

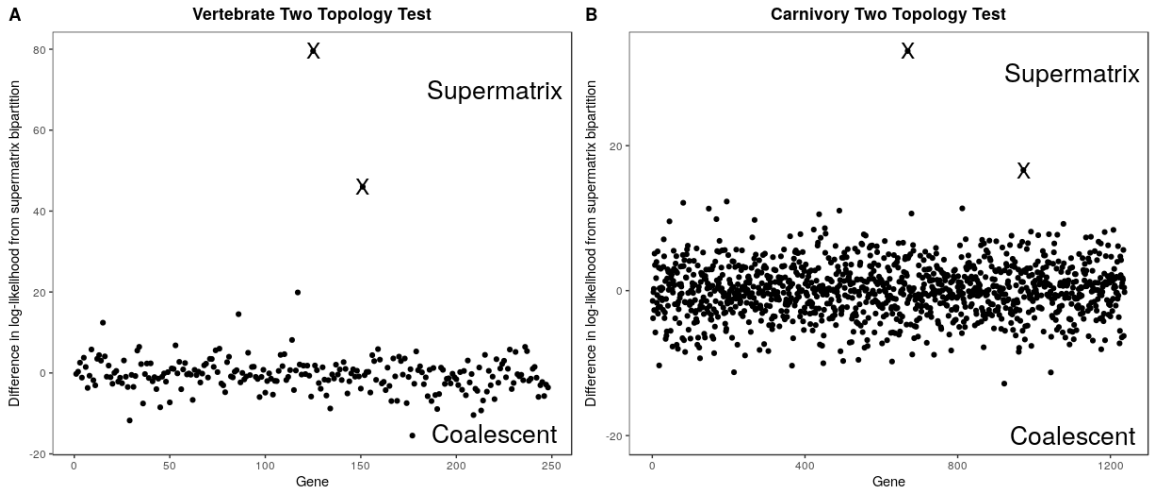


Figure 3. Identification of outlier genes using gene-wise likelihood test. A&B) Show the results of the two-topology test on the vertebrate and carnivory dataset, respectively, using the coalescent (negative values) and supermatrix (positive values) topologies as the comparison. The genes identified as outliers from the analysis are marked with an X.

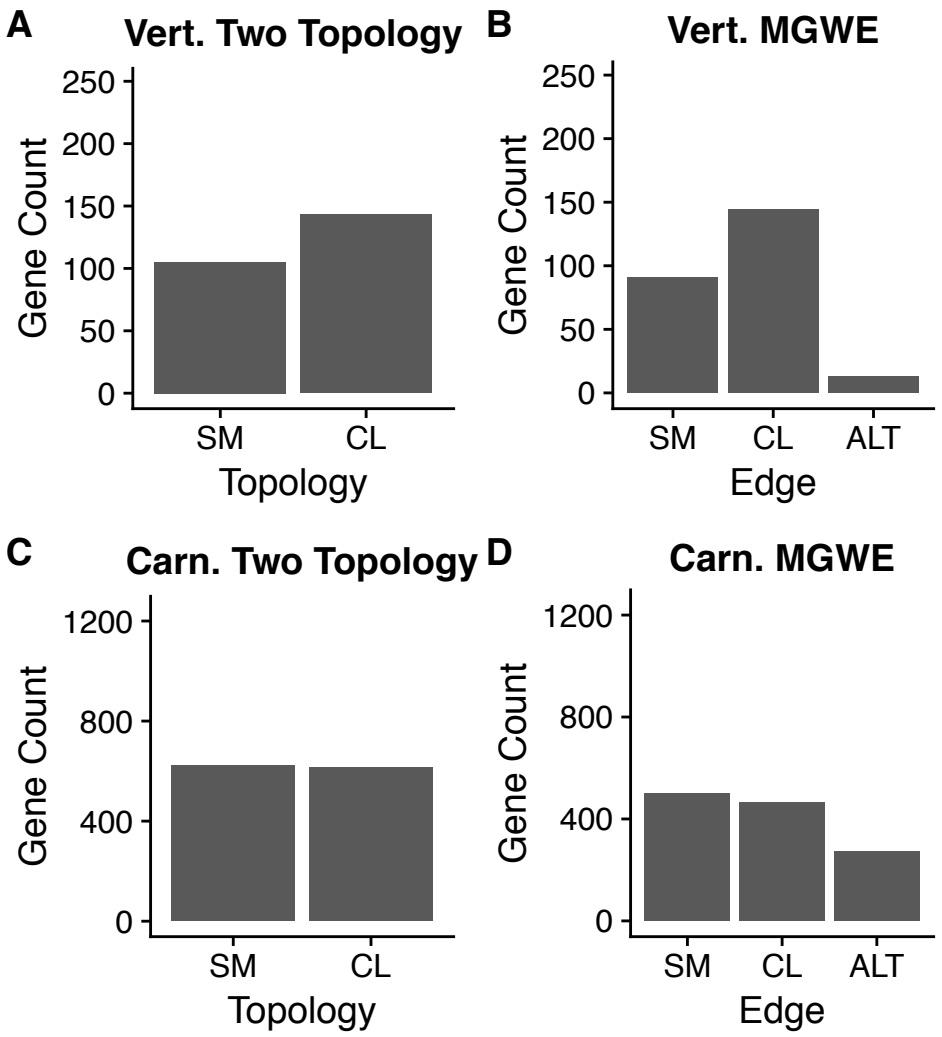


Figure 4. Bar plot representing gene counts for the two-topology and MGWE methods. A&C) represent counts of genes that support the supermatrix inferred maximum likelihood (ML) topology and the maximum quartet support species tree (MQSST), for the vertebrate and carnivory datasets respectively. B&D) Show the results of the MGWB analysis for support of the edge found in the ML analysis, the conflicting edge from the MQSST analysis, and the sum of all genes supporting an alternative conflict from an edge in the TREE SET.

559

560

561

562 **Table 1. Results of model testing the various topologies and edges.**

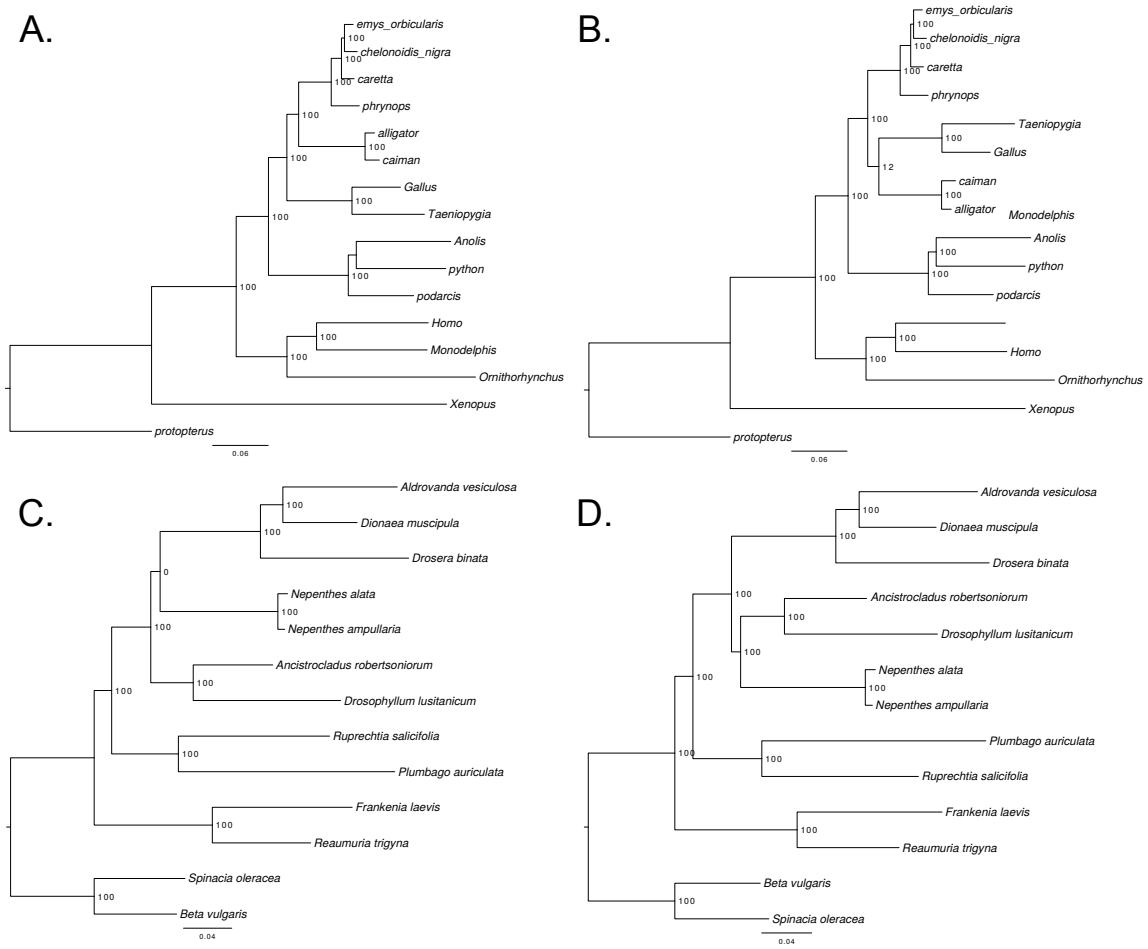
563

Vertebrate	Supermatrix	Topology	-1,047,406.05	1517	2,097,846.11	22374.01
		As Gene Trees	-1,031,489.81	6442	2,075,863.63	391.53
		Edge	-1,031,423.65	6442	2,075,731.31	259.20
	Coalescent	Topology	-1,047,410.07	1517	2,097,854.15	22382.04
		As Gene Trees	-1,031,450.71	6442	2,075,785.43	313.32
		Edge	**-1,031,294.05**	**6442**	**2,075,472.10**	**0**
	Dominant Alternative	Edge	-1,033,773.81	6442	2,080,431.62	4959.52
Carnivory	Supermatrix	Topology	-13,305,055.20	7445	26,625,000.40	36618.41
		As Gene Trees	-13,205,130.14	35873	26,595,640.58	7258.59
		Edge	-13,258,387.61	35873	26,588,521.23	139.24
	Coalescent	Topology	-13,305,130.14	7445	26,625,150.28	36768.28
		As Gene Trees	-13,262,019.55	35873	26,595,785.10	7403.10
		Edge	**-13,258,317.99**	**35873**	**26,588,381.99**	**0**
	Dominant Alternative	Edge	-13,260,106.83	35873	26,591,959.66	3577.67

564 *In the type column, "Topology" represents the supermatrix or coalescent topology with a single set
565 of branch lengths, "As Gene Trees" is the supermatrix or coalescent topology with branch lengths
566 varying among genes, and "Edge" is the MGWE analysis. The top AIC score is bolded.

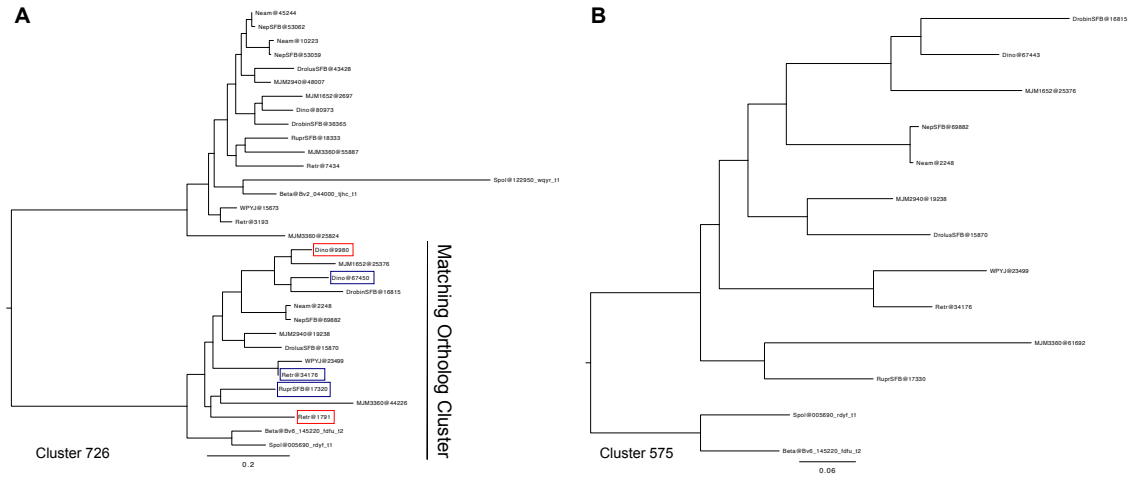
567

568



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) Species tree for vertebrate dataset inferred with all 248 genes included in the supermatrix. B) Species tree for the vertebrate dataset inferred with 8916 and 11434 removed from the supermatrix. C) carnivorous Caryophyllales species tree inferred from

all 1237 genes. D) carnivorous Caryophyllales species tree inferred with cluster575 and cluster3300 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.

Supplementary Table 1. Number of gene trees in which all the species for a given edges are present. edges correspond to node labels on Fig. 1.

Edge number	Genes containing all species for the edge
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
edge of turtles sister to birds+crocodilians	36

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

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

599 **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

600