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

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

Here we discuss an alternative to Bayes factors for identifying genes that have a disproportionate contribution to the resolution of the species tree topology. We explore gene-tree conflict by examining both site- and gene-specific log-likelihoods. Site-wise log-likelihood analyses have been employed in phylogenomic datasets previously (Castoe et al. 2009; Smith et al. 2011), primarily to compare two alternative topologies. Here, we also examine per-gene log-likelihood differences and gene-wise conflict as in Smith et al. (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 detecting overly influential genes, we ensured that maximum likelihood would recapitulate the previous species-tree results. To construct a species tree for the vertebrate dataset, the 248 individual genes vertebrate genes used in Brown and Thomson (2016) for inference of highly influential genes were concatenated pxcat (Brown, Walker and Smith, in press). The species tree was inferred with maximum likelihood as implemented in RAxML v8.2.3 (Stamatakis 2014) using the GTR+CAT model of evolution with 200 rapid bootstrap replicates performed. The use of CAT in the species tree analysis was performed to save computational time as final inference is still conducted under GAMMA. The species tree for the vertebrate dataset was inferred both with all genes present, and again inferred in the same means with the previously identified two most highly informative genes (8916 and 11434) removed (see below). The species tree inferred through maximum likelihood and containing all data from the carnivory dataset was downloaded from (XXXX). Another species tree was inferred through maximum likelihood after removing the two highly informative genes (cluster575 and cluster3300; see below) from the supermatrix.

Gene tree construction and analysis of conflict

Individual gene trees were inferred using maximum likelihood with the GTR+CAT model of evolution as implemented in RAxML. A SH-Like test (Anisimova et al. 2011), as implemented in RAxML, was performed to analyze the gene tree support. As this test examines alternative topologies by NNI, it is possible that during the test a topology with a higher likelihood is found. If a better topology was found during the test performed for this study, that topology was used in downstream analyses. All gene trees were rooted on the outgroup (*Protopterus* for the vertebrate dataset and *Beta vulgaris* and *Spinacia oleraceae* for the carnivory dataset) and any gene trees not containing the outgroup were left out of the conflict analysis. Conflict was assessed by examining taxon bipartitions as implemented in phyparts (Smith et al. 2015) with SH-Like support of < 80 treated as uninformative. The conflict was mapped on the species tree using the script phypartspiecharts.py (available from https://github.com/mossmatters/MJPythonNotebooks).

Gene log-likelihood Analysis

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

Testing for paralogy in carnivory dataset

The homolog trees created from amino acid data in the study by Walker et. al (in review) were downloaded from (XXX). We examined the maximum inclusion (Yang and Smith 2014) homologs of the amino acid data and compared the clusters containing the outlier genes to those nucleotide clusters containing the outlier genes. This allowed us to examine the possibility that the nucleotide cluster contained homology errors that would be exposed by the slower evolving amino acid dataset.

RESULTS

Likelihood based species tree inferences

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

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

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

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

We performed this same comparison of log-likelihoods between the dominant topologies on the carnivory dataset. We found two genes (cluster575 and cluster3300) that contribute disproportionately to the overall likelihood and that individually have a difference in log-likelihood scores of 33.06 and 16.63, respectively, with the average difference of log-likelihood of a gene in the supermatrix to either topology being 2.882.

Disproportionate information may potentially be a biological reality

For the carnivorous Caryophyllales dataset, we explored the possibility that the strongly conflicting genes cluster575 and cluster3300 reflected some methodological error in the assembly pipeline, as is the case for the genes identified by Brown and Thomson (2016). However, both the alignment and inferred phylogram for each gene revealed no obvious problems or potential sources of systematic error (sparse alignment, abnormally long branch lengths etc…). We also explored whether compositional heterogeneity could explain the strongly conflicting results (i.e., that the relationships were not truly conflicting, but instead incorrectly modeled). However, both RY-coding in RAxML and explicit modeling of multiple equilibrium frequencies (2, 3, or 4 composition regimes) across the tree in p4 v1.0 (Foster 2004) failed to overturn the inferred relationships. We further explored the possibility of misidentified orthology. By examining the homolog tree produced from amino acid data, we identified the ortholog from the nucleotide data to be complete (i.e., 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 Figure 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 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 among alternate topologies, high posterior probabilities were reported at every node (PP = 1.00) (Brown and Thomson 2016). The change in topology of the carnivory dataset is also remarkable, as all 1237 genes are represented in each species and the removal of two (0.0016%) of the genes resulted in a different topology with high support. Some authors have noted discrepancies between coalescent and supermatrix results in these datasets (Wickett et al. 2014; Xi et al. 2014;Walker et. al in review). The observation that a small number of genes, in the context of supermatrix analyses, can influence the resulting topology may help explain this phenomenon. The results from both datasets discussed here emphasize that even with large, high coverage datasets, supermatrix analyses may be sensitive to a small number of influential genes.

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

Brown J.M., Thomson R.C. 2016. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. syw101.

Brown J.W., Walker J.F., Smith S.A. 2017. phyx: Phylogenetic tools for Unix. Bioinformatics. accepted.

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.

Crawford N.G., Parham J.F., Sellas A.B., Faircloth B.C., Glenn T.C., Papenfuss T.J., Henderson J.B., Hansen M.H., Simison W.B. 2015. A phylogenomic analysis of turtles. Mol. Phylogenet. Evol. 83:250–257.

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

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C. V., Lovell P. V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M. V., Alfaro-Nunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield

P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jonsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alstrom P., Edwards S. V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science (80-.). 346:1320–1331.

Kobert K., Salichos L., Rokas A., Stamatakis A. 2016. Computing the internode certainty and related measures from partial gene trees. Mol. Biol. Evol. Advance Ac:1–17.

Morgan C.C., Foster P.G., Webb A.E., Pisani D., McInerney J.O., O'Connell M.J. 2013. Heterogeneous models place the root of the placental mammal phylogeny. Mol. Biol. Evol. 30:2145–56.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PLoS Biol. 14:1–24.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Moriarty Lemmon E., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.

Romiguier J., Ranwez V., Delsuc F., Galtier N., Douzery E.J.P. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol. Biol. Evol. 30:2134–44.

Ryan J.F., Pang K., Schnitzler C.E., Nguyen A.D., Moreland R.T., Simmons D.K., Koch

B.J., Francis W.R., Havlak P., Smith S.A., Putnam N.H., Haddock S.H., Dunn C.W., Wolfsberg T.G., Mullikin J.C., Martindale M.Q., Baxevanis A.D. 2013. The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. Science (80-.). 342:1242592.

Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.

Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:150.

Smith S.A., Wilson N.G., Goetz F.E., Feehery C., Andrade S.C.S., Rouse G.W., Giribet G., Dunn C.W. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature. 480:364–367.

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

Walker, J.F., Yang, Y. Moore, M.J., Mikenas, J., Brockington, S.F., Timoneda, A., Smith S.A. 2017. Widespread paleopolyploidy, gene tree conflict and recalcitrant relationships among the carnivorous caryophyllales. in review.

Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson 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., Soltis D.E., Wong G.K., Carpenter E.J., Zhang Y., Chen L., Yan Z., Xie Y., Sage R.F., Covshoff S., Hibberd J.M., Nelson M.N., Smith S.A. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Mol. Biol. Evol.:msv081-.

FIGURE CAPTIONS

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

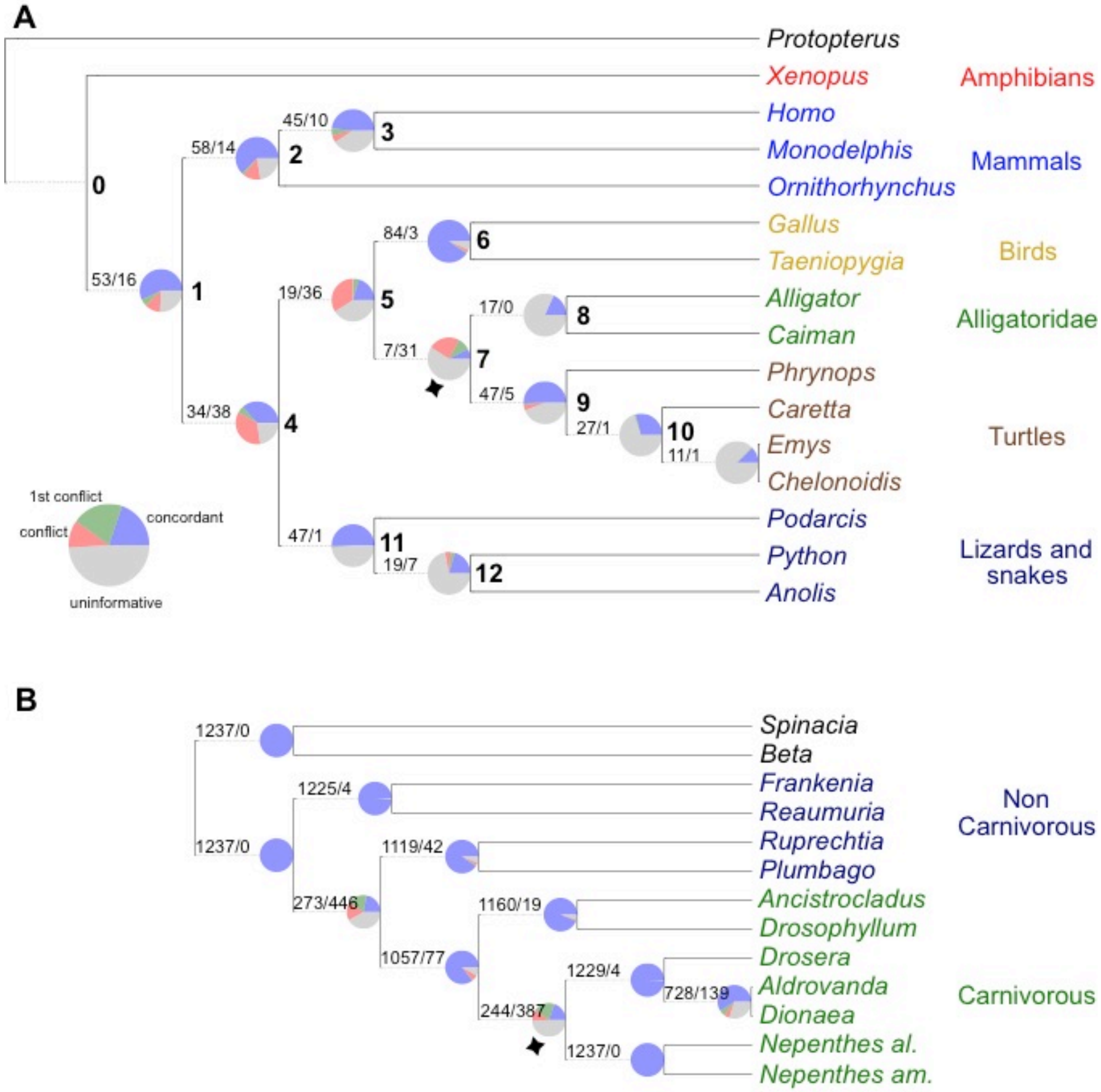

A
Protopterus
Xenopus
Amphibians
Homo
Monodelphis
Mammals
Ornithorhynchus
Gallus
Taeniopygia
Birds
Alligator
Caiman
Alligatoridae
Phrynops
Caretta
Emys
Turtles
Chelonoidis
Podarcis
Python
Lizards and snakes
Anolis
0
1
2
3
4
5
6
7
8
9
10
11
12
53/16
58/14
45/10
34/38
19/36
84/3
7/31
17/0
47/5
27/1
11/1
47/1
19/7
1st conflict
concordant
conflict
uninformative
B
Spinacia
Beta
Frankenia
Reaumuria
Non Carnivorous
Ruprechtia
Plumbago
Ancistrocladus
Drosophyllum
Drosera
Aldrovanda
Carnivorous
Dionaea
Nepenthes al.
Nepenthes am.
1237/0
1225/4
1237/0
1119/42
273/446
1160/19
1057/77
1229/4
728/139
244/387
1237/0

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

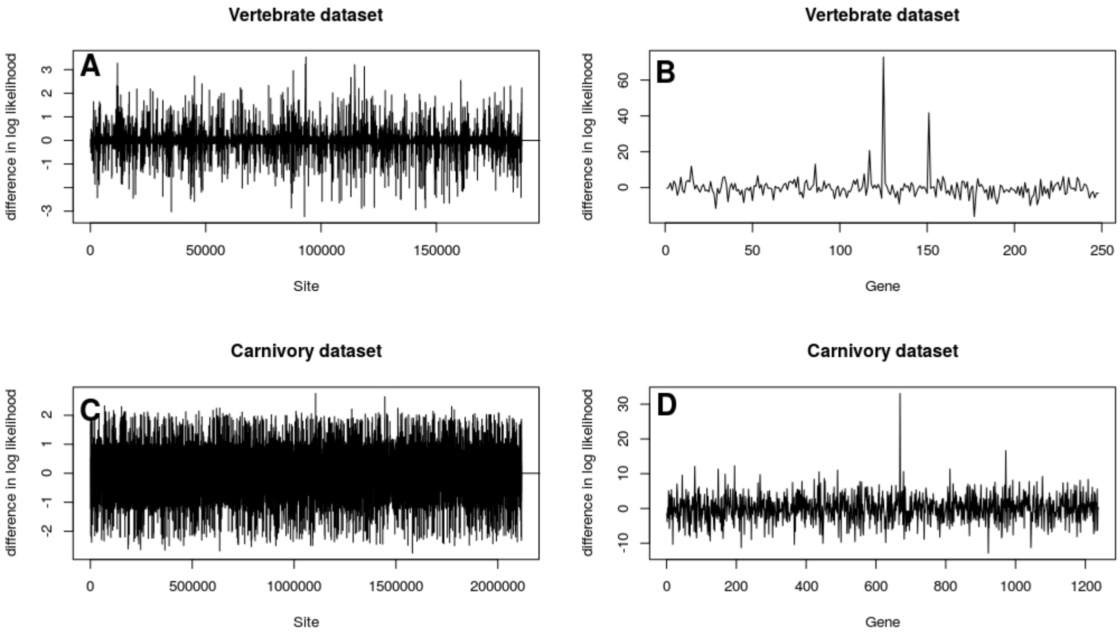
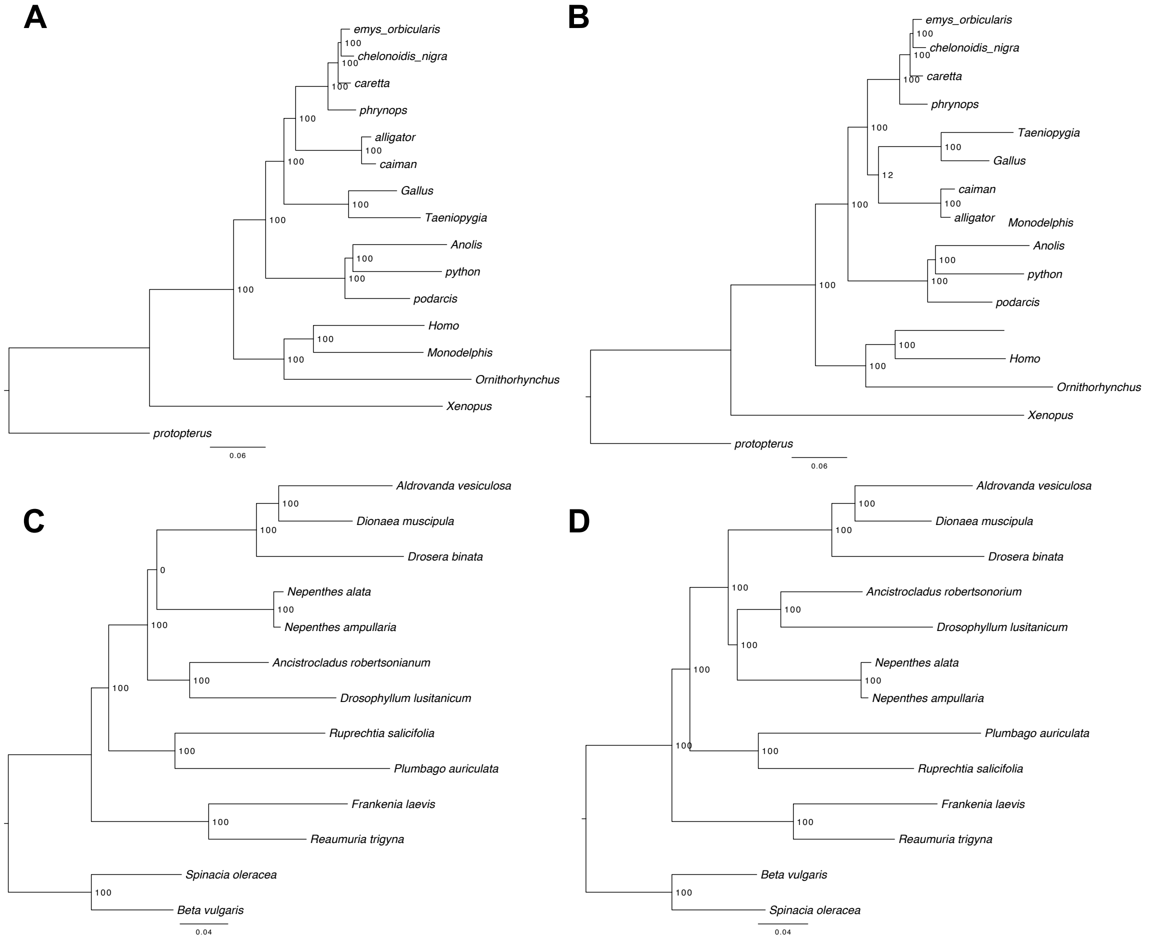


Table 1. Number of gene trees in which all the species for a given bipartition are present. Bipartitions correspond to node labels on figure 1.

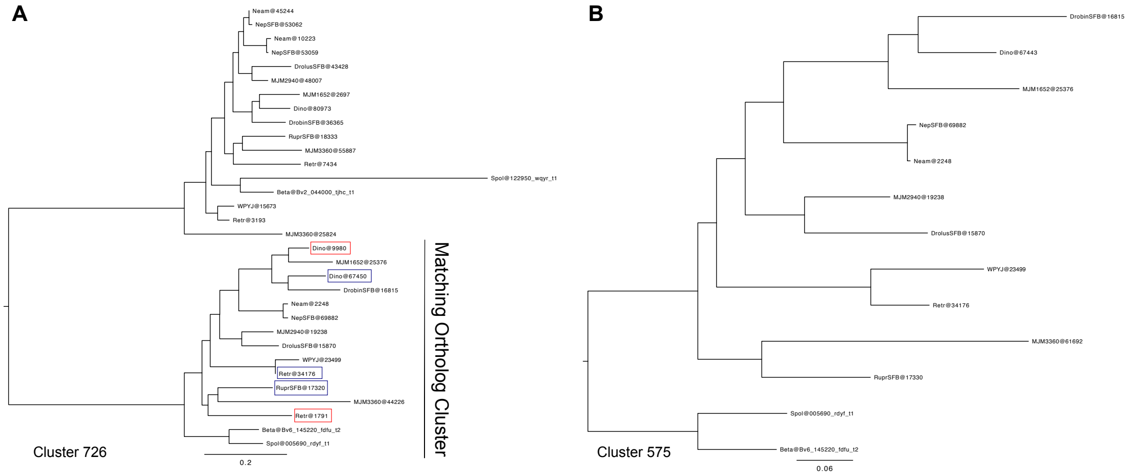
Bipartition number	Gene trees containing all species in the bipartition
0	5
1	5
2	246
3	248
4	5
5 (All turtles, Alligatoridae and birds)	6
6	248
7	6
8	23
9	36
10	45
11	69
12	51
13	94
Bipartition of turtles sister to a clade of birds and Alligatoridae	36

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.



Supplementary Table. Sources of discrepancy between the orthologs detected in highly influential nucleotide cluster575 and in matching amino acid homolog 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