Nested phylogenetic conflicts and deep phylogenomics in plants

Stephen A. Smith[1], Joseph F. Walker[1], Joseph W. Brown[2], Nathanael Walker-Hale[1]

Author Affiliation: [1]Department of Ecology and Evolution Biology, University of Michigan, [2]Department of Animal and Plant Sciences, University of Sheffield

Keywords: phylogenomics, phylogenetics, angiosperms, gene-tree conflict, genomics, coalescent

Abstract

Recent studies have demonstrated that extensive gene tree conflict underly several important phylogenetic relationships and that alternative species tree methods produce inconsistent results for recalcitrant lineages. Here, we focused on resolving several contentious, but evolutionarily significant, relationships across land plants using methods that isolate phylogenetic branches. These analyses provide insight into the source of conflict among species tree methods, disentangling aspects which may have influenced previously inferred phylogenies and providing greater confidence in estimated species relationships. Specifically, our results support the hypotheses that *Amborella* is sister to the remaining extant angiosperms, that extant gymnosperms are monophyletic, and that the Gnetales are sister to the pines. Several other contentious relationships, including the resolution of relationships among both the bryophytes and the eudicots, remain uncertain given the relatively low number of supporting gene trees. Our analyses also suggest that significant biological or systematic error may severely limit the amount of informative data. Furthermore, using a novel combinatorial heuristic, we demonstrate that the underlying conflicting signal does not support broad concatenation of gene regions, even when filtering gene regions by supporting relationships. The approach explored here offers a means to isolate and analyze underlying phylogenetic signal that can be applied across the Tree of Life.

Introduction

Over the last few years, we have come to understand that phylogenetic conflict is a common feature across the tree of life and it inhibits our ability to resolve fundamentally important relationships. Such persistent phylogenetic conflict has been noted throughout land plants, a clade which represents roughly half a million species and is one of the most diverse and ecologically significant clades on Earth. Despite the importance of land plants, several relationships which are crucial to an evolutionary understanding of key biological features remain unresolved. For example, the relationships among the lineages of bryophytes (i.e., hornworts, liverworts, and mosses) remain unclear despite extensive data collection efforts (Wickett et al. 2014; Puttick et al. 2018). One of the most heavily debated lineages in plant phylogenetics has been the placement of the monotypic *Amborella*, whose conflicting placements alter our understanding of early flowering plant evolution. *Amborella* has been variously placed as sister to Nymphaeales, as sister to all angiosperms, or as sister to the remaining Angiosperms excluding Nymphaeales (Xi et al. 2014). The resolution of *Ambrorella*, along with other contentious relationships across land plants, would provide greater confidence in our understanding of the evolution of early reproductive ecology, the evolution of floral development, and the life history of early land plants (Feild et al. 2004; Sauquet et al. 2017).

Due in large part to the recent reduction in the effort and expense required to generate molecular

sequences, researchers have amassed these large genomic and transcriptomic datasets meant to resolve fundamental phylogenetic relationships across the tree of life including in plants (Wickett et al. 2014), animals (Jarvis et al. 2014; Dunn et al. 2008; Simion et al. 2017; Whelan et al. 2017), fungi (Shen et al. 2016), and bacteria (Ahrenfeldt et al. 2017). While the goals of these data collection efforts have been to increase the overall phylogenetic support for evolutionarily significant relationships, several recent analyses have demonstrated that different datasets and analytical approaches often reconstruct strongly-supported but conflicting relationships (Feuda et al. 2017; Walker et al. 2018; Shen, Hittinger, and Rokas 2017). Underlying these conflicting results are typically strongly conflicting individual gene trees relationships (Smith et al. 2015). In some cases, one or two "outlier" genes can overrule thousands of other genes in the resolution of relationships (Shen, Hittinger, and Rokas 2017; Brown and Thomson 2016; Walker, Brown, and Smith 2018). These genes may be the result of biological processes (e.g. Walker, Brown, and Smith 2018) or systematic error (Brown and Thomson 2016) and their removal may alter the inferred species relationships. Whether dealing with overall gene tree conflict or outlier genes, the analysis of these large phylogenomic datasets requires detailed consideration.

Traditionally researchers have had two ways to deal with large phylogenomic questions: concatenated supermatrices and coalescent gene-tree / species tree methods. Supermatrix methods were, in part, developed to allow for the strongest signal to prevail when conducting phylogenetic analyses. However, it has long been understood that the 'total evidence' paradigm (Kluge 1989), where the true history will 'win out' if only enough data are collected, is untenable. For example, genes with real and conflicting histories are present within datasets (Maddison 1997). Later, a new paradigm was heralded for phylogenetic systematics (Edwards, Liu, and Pearl 2007; Liu et al. 2009): that of 'species tree' inference, where the strict assumption that gene trees must share the same topology is relaxed (Edwards 2009; Edwards et al. 2016). For phylogenomic studies, analyses that accommodate incomplete lineage sorting (ILS) are often conducted alongside analyses that concatenate genes into a supermatrix, with almost all studies resulting in discordance involving at least one contentious focal relationship. Despite the wide adoption of both approaches, concatenation and species tree methods make different assumptions. For example, concatenation approaches, while allowing mixed molecular models and gene-specific branch lengths, assume a single underlying tree. Coalescent approaches, depending on the implementation, may assume that all conflict is the result of ILS, that all genes evolved under selective neutrality and constant effective population size, that all genes contain enough information to properly resolve nodes, and that gene trees are estimated accurately (Springer and Gatesy 2016).

It may be the case that neither of these two approaches is valid for unfiltered phylogenomic datasets given the underlying variation and the diversity of processes leading to gene tree discordance. Importantly, the suitability of each method may differ widely clade-to-clade based on which biological processes have occured during evolutionary history. Some researchers have explored other approaches that allow for incorporation of the processes that lead to gene tree discordance (Ané et al. 2006; Boussau et al. 2013). However, these two widely-used methods are often computationally intractable for the enormous scale of current genomic datasets. Indeed, the distinction between concatentation and coalescent-based methods, and their conclusions for certain contentious relationships, is such that systematists are seemingly faced with a dichotomy. We argue that methods and approaches which focus on analyzing a given contentious relationship in the data render this dichotomy a false one, and should be pursued by additional methods to advance our understanding of the Tree of Life.

Here, we reanalyzed a large plant genomic dataset (Wickett et al. 2014) to isolate phylogenetic signal of particularly contentious relationships. Specifically, we explored an alternative to concatenation

and species-tree approaches for analyzing the signal for individual species relationships. We examine systematic error, nested conflicting relationships, and quantify the extent of gene tree disagreement. Furthermore, we investigated the assumption of a single underlying tree, by examining the suitability of a concatenation approach for species tree resolution. By taking this broad information-centric approach, we hope to shed more light on the evolution of plants and present a more biologically-informed method with broad applicability for phylogenomic datasets across the Tree of Life.

Results and Discussion

Conflict analyses

We conducted analyses comparing gene trees to each other and to the maximum likelihood tree (Fig. 1) based on the concatenated maximum likelihood (ML) analysis from Wickett et al. (2014). We found that both gene tree conflict and support varied through time with support increasing toward the present (Fig. 2). We aimed to resolve specific contentious relationships, the resolution of which has either been debated in the literature or been considered important in resolving key evolutionary questions, to the best of the ability of the underlying data (Table 1).

Several conflicting relationships were the result of systematic error in the underlying data. In order to minimize the impact of systematic error on the estimation of relationships, we excluded obvious error where possible. For example, we found 258 of 852 gene trees contained non-land plant taxa that fell within the land plants. While these errors may not impact the estimation of relationships within eudicots, they will impact the estimation of relationships at the origin of land plants. Therefore, we excluded gene trees for which there was not previously well established monophyly of the focal taxa (i.e., involving the relationship of interest). We also identified 68 gene trees that possessed very long estimated branch lengths (> 2.5 expected substitutions per site). We conservatively considered these to contain potential errors in homology (Yang and Smith 2014). While these genes demonstrate patterns associated with systematic error, they also, likely, contain information for several relationships. However, some error may be the result of misidentified orthology that will mislead estimation of phylogenetic relationships, even if this error may not impact all relationships inferred by the gene. Therefore, to minimize sources of systematic error, we took a conservative approach and excluded these genes from additional analyses.

We found several contentious relationships display patterns similar to those expected under an ILS model, such as at the origin of angiosperms (e.g., *Amborella* in Table 1), where the number of genes supporting alternative resolutions were roughly equal. This corresponds to the recovery of these relationships by coalescent analyses in the original study (Wickett et al. 2014). However, in addition the the number of genes, we also compared the sum of the difference in the likelihoods for relationships for each gene (see Material and Methods). The difference between the number of gene trees supporting relationships and the difference in the summed likelihoods provide insight into the reason for discordance between concatenated ML analyses and coalescent analyses. For example, the relationship involving Gnetales and the conifers as sister (Gnetifers) was recovered in coalescent-based analysis and is supported by more genes (Table 1). However, the sum of the differences in the log-likelihoods of alternative resolutions support the Gnepine relationship (i.e., Gnetales sister to Pinales), as found in the ML analyses. The gene trees equivocally support several relationships for eudicots and bryophytes. However, once log-likelihoods are compared, a dominant relationship emerged (Fig 2).

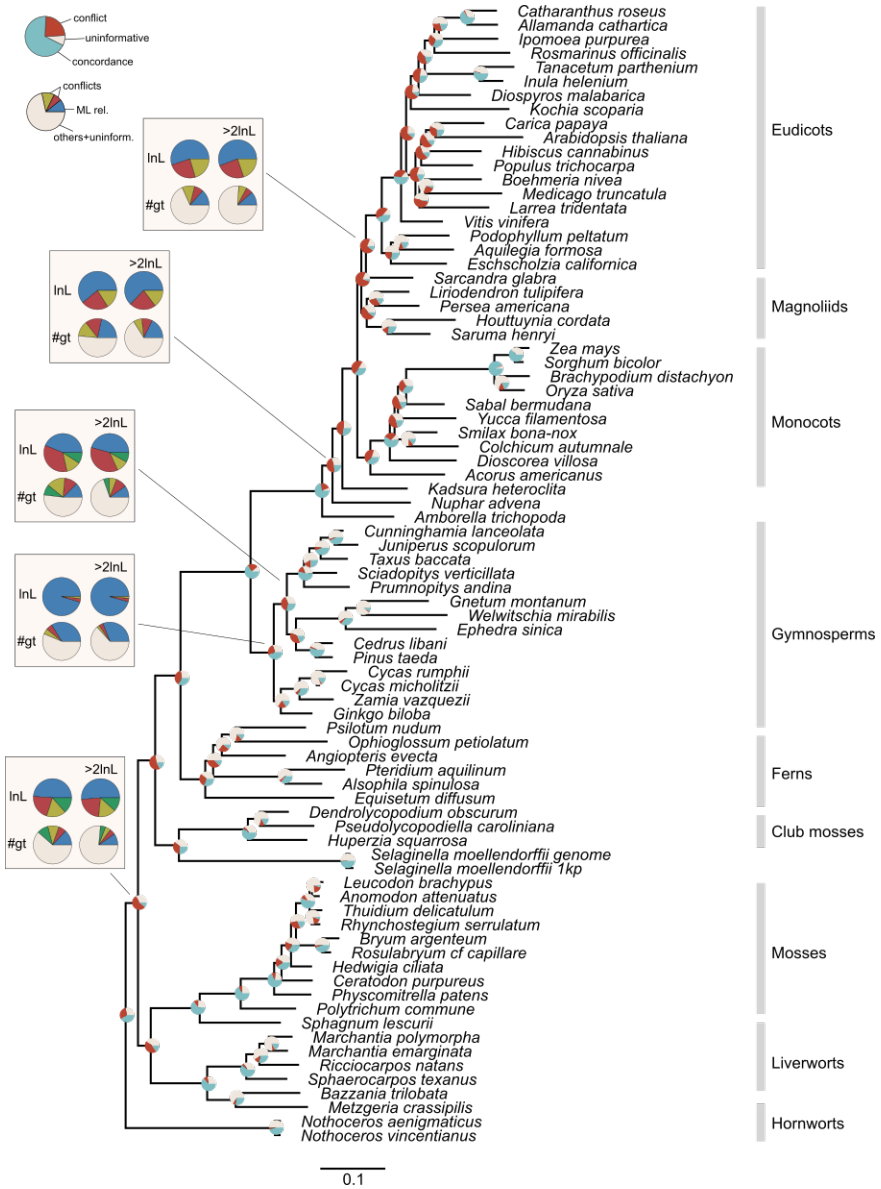

Figure 1: Phylogeny of land plants with pie charts at nodes illustrating conflict, concordance, and informativeness of the gene tree set without any filtering. Inset boxes show summed differences in log likelihoods (top row) and the number of gene trees (bottom row) that support the relationship shown in the tree and the dominant conflicting relationships. Right pie charts in the inset box show results when only differences greater than 2 log likelihoods are considered. See also Table 1.

*Table 1. Comparison of the number of genes and the difference in the likelihood (DlnL) with relationships ordered based on support. * indicates relationships present in the ML tree.*

Major clade	Resolutions	Genes	Genes (> 2lnL)	DlnL	DlnL > 2
Bryophytes	Hornworts sister*	110	83	677.6	654.1
	Liverworts sister	56	41	294.1	280.8
	Mosses+liverworts	81	40	228.9	190.2
	All monophyly	81	37	185.3	148.5
Gymnosperms	monophyly*	288	264	7259.0	7233.8
	Gnetum sister	45	31	229.8	216.0
	Cycas sister	39	18	120.3	105.2
Gymno relat.	Gnepine*	107	85	1017.2	994.4
	conifers	93	79	800.0	787.2
	Gnetifers	134	55	288.1	217.8
	Gnetales sister	76	40	211.2	176.3
Amborella	*Amborella* sister*	184	152	1501.1	1470.0
	Amborella+*Nuphar*	118	75	564.2	526.3
	Nuphar sister	111	62	392.2	345.2
Eudicots	Magnoliids+eudicots*	114	98	1223.4	1204.3
	Monocots+eudicots	66	49	541.5	526.5
	Monocots+magnoliids	90	58	453.3	425.5

Nested analyses

Given the variation in support and conflict through time (Fig. 2), many genes that contain signal for a particular relationship may disagree with the resolution at other nodes. To examine these patterns of nested conflict, we examined the genes that support the resolution of the eudicot relationships (Fig 3). In a set of 127 genes which supported the eudicot relationships recovered in the original ML analysis, 98 survived filtering for outgroup placement, branch length, and support with a statistically significant difference in lnL (> 2 ; Edwards 1984). 63 of these genes supported the monophyly of gymnosperms, and among those 63 only 25 supported a sister relationship between pines and *Gnetum*.

This analysis demonstrates the significant variation in the support for different relationships throughout the tree. Even without gene tree conflict, it is perhaps naïve to expect a single gene to have high support throughout a large part of the tree of life (see Penny et al. (1990); MUTOG: the 'Myth of a Universal Tree from One Gene'). This is especially unlikely when the phylogeny of interest spans relatively old and young ages as is the case explored here. For this reason, some researchers have thus argued that concatenating genes effectively combines data informative at various scales and so provides the necessary information to better resolve deep and shallow nodes (e.g., Mirarab, Bayzid, et al. 2014). However, it is not clear whether conflicting signal can be overcome with concatenation, and so we address this question below.

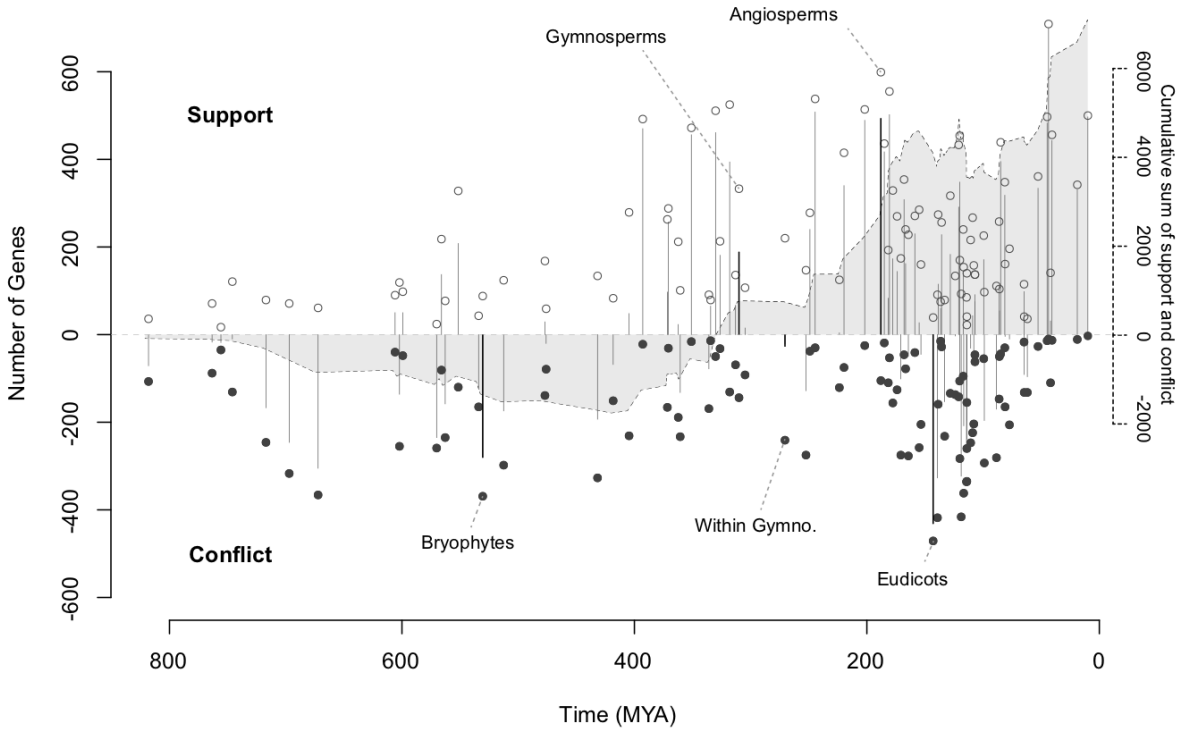

Figure 2: Examination of support and conflict in relation to time across all nodes with time as estimated using TIMETREE (Hedges, Dudley, and Kumar 2006). The differences between support and conflict are noted with vertical lines. The cumulative sum of support and conflict through time is noted in solid grey. Focal nodes from Fig. 1 are identified.

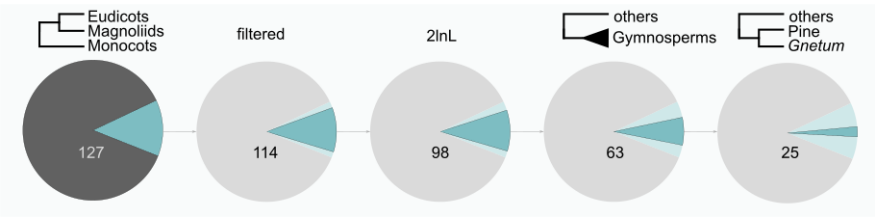

Figure 3: Nested patterns of support with genes associated with the resolution of eudicots. From left to right are shown the genes that support eudicots as sister to magnoliids (far left), those genes filtered as not having any outgroup errors or long branch lengths, those genes that support the resolution by at least 2lnL, those genes that support monophyletic gymnosperms, and finally those genes that support the Gnepine relationship.

Combinability of genes

Despite the potential benefits of concatenating genes (i.e., amplifying weak phylogenetic signal), the underlying model of evolution for a concatenated analysis assumes topological concordance among gene tree histories. Given extensive gene conflicts, nested and otherwise, it may be that these assumptions should often be violated. Whether genes should be combined for a concatenated analysis has been discussed (Huelsenbeck, Bull, and Cunningham 1996; Theobald 2010; Walker, Brown, and Smith 2018) along with the recent development of Bayesian methods meant to address these issues (Neupane et al. 2018). Here, given the large size of the dataset, we rely on information theoretic methods (e.g., AICc) that greedily test combinability of genes sets based on Robinson Foulds distances to examine whether genes can be justifiably concatenated despite heterogeneity in information content throughout the phylogeny. We refer to this as the COMBination of datasets (COMB) method. Because our approach bears conceptual similarity to algorithms used to estimate the optimal partitioning scheme (e.g. PartitionFinder, Lanfear et al. 2016), we compared combinable subsets to those recommended by the implementation of the PartitionFinder algorithm in IQ-TREE (Kalyaanamoorthy et al. 2017, referred to as MERGE here). Since an exhaustive search of the entire dataset is intractable, we examined the combinability of those genes that support the eudicot lineages to be sister to the magnoliid lineages (Fig. 3). We conducted analyses of two sets of genes: those that support the relationship with greater than 2 lnL versus alternative relationships (98 genes; 'CombinedSet'), and those that display the relationship in the ML gene tree and have SH-aLRT support greater than 80 (44 genes; 'MLSet'). These two sets were chosen because the first set was already examined as part of this study and the second is a typical cutoff used in standard systematics analyses (Guindon et al. 2010).

No method or gene set supported the concatenation of all genes that supported the focal eudicot relationship (see Table 2). The COMB method on the 'CombinedSet' supported some concatenation of 29 of 98 total genes: 13 sets consisting of two genes, and one set which consisting of three genes. The MERGE method supported concatenation of 28 of 98 genes in the Combined Set: 8 sets of two genes, and four sets of three genes (see Table 2 for more details). Despite similarity in the number of genes to be concatenated, the COMB and MERGE results did not contain any identical concatenated sets. We constructed phylogenies of each concatenated set and compared the inferred topologies (results in Table 2). Despite filtering on the magnoliids as sister to eudicots relationship, not all concatenated sets recovered this relationship with greater than 80 SH-aLRT.

Concatenation is a common means for analyzing large phylogenomic analyses, and so it may be surprising that a relatively small number of genes support concatenation. However, this may be expected considering the extensive gene tree conflict. While concatenation may be helpful for exploratory inference to identify dominant signal, it may not be the best approach to address specific and contentious relationships. Further analysis, such as the one described here, into nodes that conflict with species-tree methods or are surrounded by gene tree conflict, should be pursued to uncover the most robust phylogenetic hypothesis upon which to base other evolutionary hypotheses.

Table 2. Comparison of partitioned subsets between combining strategies

Algorithm	Gene set	Genes	Sets	Partitioned Topology	Subset Relationships
MERGE	combined	98	12 (4x3, 8x2)	magnoliids+eudicots (100)	magnoliids+eudicots (50%)

Algorithm	Gene set	Genes	Sets	Partitioned Topology	Subset Relationships
					magnoliids+monocots (0)
					monocots+eudicots (0)
	ML	44	9 (1x3, 8x2)	magnoliids+eudicots (100)	magnoliids+eudicots (67%)
					magnoliids+monocots (0)
					monocots+eudicots (0)
COMB	combined	98	14 (1x3, 13x2)	magnoliids+eudicots (100)	magnoliids+eudicots(50%)
					magnoliids+monocots (0)
					monocots+eudicots (0)
	ML	44	10 (3x3, 7x2)	magnoliids+eudicots (100)	magnoliids+eudicots (90%)
					magnoliids+monocots (0)
					monocots+eudicots (0)

Brackets following a partitioned topology give the SH-aLRT score for that branch, while percentages following a subset relationship give the proportion of individual partition gene trees supporting the specified relationship with ≥ 80 SH-aLRT

Implications for plant phylogenetics

The results presented here provide strong support for several relationships that have long been considered contentious, and indicate probable resolutions for others. For example, we found more genes and higher likelihoods for 1) *Amborella* being sister to the rest of angiosperms and 2) that gymnosperms are monophyletic. Several relationships (e.g., among the eudicots and relatives as well as the hornworts, liverworts, and mosses) lack enough information to confidently accept any of the alternative resolutions. Rather than being dismayed at this apparent failure, we regard this lack of signal as extremely valuable information, as it informs where future effort should be focused. Though we identified the relationship that was more strongly supported by the data (Table 1), the differences between the alternatives were so slight that the current dataset is likely unable to confidently resolve this debate and conducting additional analyses with expanded taxa and gene regions is warranted.

Among the strongly supported hypotheses, the placement of *Amborella* continues to be a point of major contention within the plxant community. *Amborella* is a tropical tree with relatively small flowers, while the Nymphaeales are aquatic plants with relatively large flowers. The resolution of these taxa in relation to the remainder of the flowering plants will inform the life history or early angiosperms (Dark and Disturbed, Feild et al. (2004)) as well as the lability of life history and floral traits. Our results suggest *Amborella* is sister to all other extant angiosperms, and implies that rates of evolution need

not be particularly fast in order to understand the morphological differences between a tropical tree (*Amborella*) and water lilies (Nymphaeales). Strong support for the monophyly of gymnosperms implies that the disparity of extant Gymnosperm taxa, including the morphologically diverse Gnetales, emerged post-divergence with the angiosperm lineage. This reinforces analyses of LEAFY homologs, which recover Gymnosperm paralogs as monophyletic groups (Sayou et al. 2014), and also lends support to shared characteristics between Gnetales and angiosperms resulting from convergent evolution (Bowe, Coat, and others 2000; Hansen et al. 1999).

For contentious relationships only weakly supported here, there are several biological questions that will be answered once these are confidently resolved. The data and analyses presented here suggest that hornworts are sister to all other land plants. This is consistent with some studies (Nickrent et al. 2000; Nishiyama and Kato 1999), but contradicts the results of others (Cox et al. 2014; Karol et al. 2010; Qiu et al. 2006), including some but not all results of a recent re-analysis of this dataset (Puttick et al. 2018). If the position of hornworts presented here holds with additional data, it implies that the absence of stomata in liverworts and some mosses is a derived state resulting from loss of the trait, suggests a single loss of pyrenoids in non-hornwort land plants (but see Villarreal and Renner 2012), and questions some inferences on the characteristics of hornwort sporophytes (Qiu et al. 2006). Among gymnosperms, these data suggest that Gnetales are sister to pines (the "Gnepine" hypothesis; Chaw et al. 2000), further supporting the lability and rapid evolution of morphological disparity within the group. Finally, magnoliids are inferred as sister to the eudicot lineages, which has implications on the origin and divergence times of eudicots and monocots.

Implications for future phylogenomic studies

A panacea does not currently exist for phylogenomic analyses. In part, this may be the result of methods meant to serve too many functions, or applied to use-cases beyond their original design. Some researchers aim to determine the relative support for contentious relationships. Others only wish to construct a reasonable, if not ideal, phylogeny for downstream analyses. Others still may be primarily interested in gene trees. Researchers seeking to perform large-scale phylogenetic inference typically use quartet-based species tree approaches, and/or concatenation (while modelling some non-topology related gene-specific properties). The underlying conflict identified by many researchers (Wickett et al. 2014; Puttick et al. 2018) suggests that concatenation, while helpful for identifying the dominant signal, may not be ideal for addressing contentious nodes. Our analyses allowed for the examination of contentious nodes while accommodating for gene tree heterogeneity without the requirement for concatenation. Furthermore, our targeted exploration of the combinability of gene regions found that very few genes are optimally modelled by concatenation, even when filtering on those genes that support one contentious relationship. While concatenation may be a relatively fast method for analyzing extremely large datasets, it may not be strictly appropriate for both statistical and biological reasons and may not be helpful for addressing difficult-to-resolve phylogenetic hypotheses.

The most common alternative to concatenation, coalescent species tree approaches, often accomodate one major source of conflict in gene trees without concatenation, ILS (Mirarab, Reaz, et al. 2014). However, the most sophisticated model-based coalescent approaches are often not computationally tractable for phylogenomic analyses because of the large sizes of the datasets (Ané et al. 2006; Boussau et al. 2013). Instead, most phylogenomic analyses that accommodate ILS use quartet methods (e.g., ASTRAL) that, while fast and effective, do not account for multiple sources of conflict and make several

other assumptions that may or may not be reasonable given the dataset (e.g. equal weighting of gene trees regardless of properties of the underlying genes). Some researchers have suggested that a solution may be to filter the data to include only those genes that conflict due to ILS (Knowles et al. 2018; Huang et al. 2017) or that agree with the accepted relationships (Doyle et al. 2015; Smith, Brown, and Walker 2018). However, for datasets with a broad scope, several processes may be at play throughout the phylogeny and it may not be possible to filter based on a single underlying process.

Here, we argue that, to address support for contentious relationships, focused branch-based analyses can provide a thorough examination of the influence on phylogenetic inference from the underlying data. With the explosion of genomic resources from new projects such as 10KP (Cheng et al. 2018), computationally efficient methods focused on specific contentious relationships will be necessary to approach challenges inherent in large datasets.

Conclusions

The results presented here provide strongly supported resolutions for two contentious relationships that have been hotly debated in the literature: that *Amborella* is the sister lineage to all other angiosperms, and that gymnosperms are monophyletic. These results have significant implications for understanding the evolution of land plants and the nature of the ancestral angiosperm. We find weak support for other contentious relationships, and suggest that these should be revisited once other datasets are amassed.

Despite the ability of the methods explored here to accomodate the underlying gene tree uncertainty, the results presented here rely on the information available in the underlying dataset. While this dateset is not comprehensive, it *does* represent extensive sequencing of transcriptomes and genomes for the taxa included. We can say, with confidence, what these data support or do not support, but different datasets (e.g., based on different taxa, different homology analyses) may have stronger signal for relationships that are resolved more equivocally here. We recommend analyzing these future datasets with an eye toward hypotheses of specific phylogenetic relationships. Our novel approach provides insight into several of the most contentious relationships across land plants and is broadly applicable among different groups. Approaches that ascertain the support for alternative resolutions should be used to resolve contentious branches across the Tree of Life.

Materials and Methods

Datasets

We acquired and analyzed the Wickett et al. (2014) dataset of transcriptomes and genomes covering plants available from http://mirrors.iplantcollaborative.org/onekp_pilot. There were several different filtering methods and approaches used in the original manuscript and, based on conversations with the corresponding author, we analyzed the filtered nucleotide dataset with third positions removed. The third positions were removed because of the problems with variation and GC content that causes problems with the placement of the lycophytes (Wickett et al. 2014). This dataset consisted of 852 aligned genes. We did not conduct any other filtering or alteration of these data before conducting the analyses performed as part of this study.

Phylogenetic analyses

We calcuated gene trees for each of the 852 genes using iqtree (v. 1.6.3; Nguyen et al. 2014). We used the GTR+G model of evolution and calculated maximum likelihood trees along with SH-aLRT values (Guindon et al. 2010). For all constrained analyses, we conducted additional maximum likelihood analyses with the same model of evolution but constrained on the branch of interest.

Conflict analyses

We conducted several different conflict analyses. First, we identified the conflicting branches between the maximum likelihood gene trees, ignoring branches that had less than 80% SH-aLRT (Guindon et al. 2010), and the maximum likelihood tree from the original publication (Fig. 2; Wickett et al. 2014). These analyses were conducted using the program `bp` available from `https://github.com/FePhyFoFum/gophy`. We reported the conflicting and concordant gene trees (Fig. 1). We placed these conflicting and supporting statistics in a temporal context by calculating the divergence times of each split based on the TIMETREE of life (Hedges, Dudley, and Kumar 2006). By examining the dominant conflicting alternatives, we established which constraints to construct and compare for further analyses. Because the gene regions contain partially overlapping taxa, automated discovery of all conflicting relationships concurrently can be challenging. To overcome these challenges, we examine each constraint individually.

To determine the difference in the lnL values among conflicting resolutions, we conducted the constrained phylogenetic analyses (with parameters described in the *Phylogenetic analyses* section above) and compared the lnL values of the alternative resolutions. We then examined those results that had a difference in the lnL of greater than 2, as is considered standard for statistical significance (Edwards 1984). For each gene, we noted the relationship with the highest log-likelihood and summed the difference of that and the second best relationship ($DlnL$) across all genes.

We also examined nested conflicts. In particular, for the genes identified as supporting the dominant relationship of the eudicot lineages, we examined the distribution of conflict. We then examined those genes that supported both the eudicot lineages and the relationship of *Amborella* as sister to the rest of angiosperms. Finally, of those genes, we determined which supported the alternative Gymnosperm rleationships. We conducted each of these nested analyses using the same methods as described above.

Concatenation tests

To explore whether concatenation was supported for different sets of genes, we conducted model fit analyses on subsets of the data. We concatenated the data using the phyx prgram pxcat (Brown, Walker, and Smith 2017) and we calculated and compared Aikaike Information Criterion scores that were corrected for sample size (AICc) on concatenated and unconcatenated analyses. We used the number of sites in an alignment as the sample size for the AIC correction as also calculated by iqtree. An AIC framework has been commonly used extensively in molecular model comparisons and has been used by several authors (e.g., Kubatko 2009; Theobald 2010; Walker, Brown, and Smith 2018) for comparisons between phylogenies and datasets. Because conducting every possible comparison of every possible combination of genes is unfeasible, we instead constructed graphs based on Robinson Foulds (RF; Robinson and Foulds 1981) distance without considering branch lengths. As above, we ignored branches with less than 80 SH-aLRT (Shimodaira and Hasegawa 1999) support as calculated

by IQ-TREE. For RF comparisons where taxa were partially overlapping, we removed tips that were present in only one tree before the comparison was calculated. These graphs describe a distance between genes based on topology and so we sorted the RF distances by the shortest and compared concatenation vs separate gene trees. If combined analyses resulted in a lower AIC score, any future comparison involving any of the constituent genes of the combination considered the combined gene set (and not the invdividual gene). For example, if gene 1 and 2 were combined based on AIC score, then an attempt to combine gene 3 and 1 would consider combining 3 and 1+2 and not 3 and 1. This allowed for concatenated datasets to grow to more than 2 genes. This analysis was effectively a greedy hill climbing analysis. To demonstrate the effectiveness of the approach, we conducted small tests (results in the Supplementary Materials).

As these analyses were conducted for demonstration purposes, we did not conduct exhaustive testing of combinability of the entire dataset. Instead, we conducted these tests on two gene sets that supported the eudicot relationship. First, we tested the set of genes that supported the eudicot relationship in the ML tree that did not have a branch length longer than 2.5 and did not have outgroup taxa falling in the ingroup. Second, we tested the set of genes that did not only support the relationship in the ML tree but also displayed the relationship in the ML gene tree with SH-aLRT support higher than 80 and with no outlying branch lengths or outgroup taxa falling in the ingroup.

Concatenated analyses were conducted using `iqtree v. 1.6.3` and the `-sp` option for branch lengths unlinked among partitions. We also tested the `-spp` and `q` options for proportional branches and shared branch lengths respectively. However, these resulted in fewer concatenated branches. A more thorough examination of these options and their behaviour is the focus of future studies.

We compared the results of our analyses to the PartitionFinder 'greedy' algorithm implemented in IQ-TREE using the option `-m MERGE`, specifying the GTR+G model and assessing partitions with the branch-unlinked model with `-sp`. We compared the gene trees of each merged partition in IQ-TREE with `-sp` and `-m GTR+G` and assessed the optimal partitioning scheme on the full data similarly with `-sp` and `-m GTR+G`.

Acknowledgments

This work was supported by funding from NSF DEB 1354048 (J.F.W. and S.A.S.) and NSF AVATOL 1207915 (J.W.B. and S.A.S.). We appreciate comments from Ning Wang, Caroline Parins-Fukuchi, Diego Alvarado Serrano, Greg Stull, Drew Larson, Hector Fox, and Richie Hodel.

References

Ahrenfeldt, Johanne, Carina Skaarup, Henrik Hasman, Anders Gorm Pedersen, Frank Møller Aarestrup, and Ole Lund. 2017. "Bacterial Whole Genome-Based Phylogeny: Construction of a New Benchmarking Dataset and Assessment of Some Existing Methods." *BMC Genomics* 18 (1). BioMed Central: 19.

Ané, Cécile, Bret Larget, David A Baum, Stacey D Smith, and Antonis Rokas. 2006. "Bayesian Estimation of Concordance Among Gene Trees." *Molecular Biology and Evolution* 24 (2). Oxford University Press: 412–26.

Boussau, Bastien, Gergely J Szöllősi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. 2013. "Genome-Scale Coestimation of Species and Gene Trees." *Genome Research* 23 (2). Cold Spring

Harbor Lab: 323–30.

Bowe, L Michelle, Gwénaële Coat, and others. 2000. "Phylogeny of Seed Plants Based on All Three Genomic Compartments: Extant Gymnosperms Are Monophyletic and Gnetales' Closest Relatives Are Conifers." *Proceedings of the National Academy of Sciences* 97 (8). National Acad Sciences: 4092–7.

Brown, Jeremy M, and Robert C Thomson. 2016. "Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses." *Systematic Biology* 66 (4). Oxford University Press: 517–30.

Brown, Joseph W, Joseph F Walker, and Stephen A Smith. 2017. "Phyx: Phylogenetic Tools for Unix." *Bioinformatics* 33 (12). Oxford University Press: 1886–8.

Chaw, Shu-Miaw, Christopher L Parkinson, Yuchang Cheng, Thomas M Vincent, and Jeffrey D Palmer. 2000. "Seed Plant Phylogeny Inferred from All Three Plant Genomes: Monophyly of Extant Gymnosperms and Origin of Gnetales from Conifers." *Proceedings of the National Academy of Sciences* 97 (8). National Acad Sciences: 4086–91.

Cheng, Shifeng, Michael Melkonian, Stephen A Smith, Samuel Brockington, John M Archibald, Pierre-Marc Delaux, Fay-Wei Li, et al. 2018. "10KP: A Phylodiverse Genome Sequencing Plan." *GigaScience* 7 (3). Oxford University Press: giy013.

Cox, Cymon J., Blaise Li, Peter G. Foster, T. Martin Embley, and Peter Civáň. 2014. "Conflicting Phylogenies for Early Land Plants Are Caused by Composition Biases Among Synonymous Substitutions." *Systematic Biology* 63 (2). Oxford University Press: 272–79. http://dx.doi.org/10.1093/sysbio/syt109.

Doyle, Vinson P, Randee E Young, Gavin JP Naylor, and Jeremy M Brown. 2015. "Can We Identify Genes with Increased Phylogenetic Reliability?" *Systematic Biology* 64 (5). Oxford University Press: 824–37.

Dunn, Casey W, Andreas Hejnol, David Q Matus, Kevin Pang, William E Browne, Stephen A Smith, Elaine Seaver, et al. 2008. "Broad Phylogenomic Sampling Improves Resolution of the Animal Tree of Life." *Nature* 452 (7188). Nature Publishing Group: 745.

Edwards, Anthony William Fairbank. 1984. *Likelihood*. CUP Archive.

Edwards, Scott V. 2009. "Is a New and General Theory of Molecular Systematics Emerging?" *Evolution* 63 (1). Wiley Online Library: 1–19.

Edwards, Scott V, Liang Liu, and Dennis K Pearl. 2007. "High-Resolution Species Trees Without Concatenation." *Proceedings of the National Academy of Sciences* 104 (14). National Acad Sciences: 5936–41.

Edwards, Scott V, Zhenxiang Xi, Axel Janke, Brant C Faircloth, John E McCormack, Travis C Glenn, Bojian Zhong, et al. 2016. "Implementing and Testing the Multispecies Coalescent Model: A Valuable Paradigm for Phylogenomics." *Molecular Phylogenetics and Evolution* 94. Elsevier: 447–62.

Feild, Taylor S, Nan Crystal Arens, James A Doyle, Todd E Dawson, and Michael J Donoghue. 2004. "Dark and Disturbed: A New Image of Early Angiosperm Ecology." *Paleobiology* 30 (1). BioOne: 82–107.

Feuda, Roberto, Martin Dohrmann, Walker Pett, Hervé Philippe, Omar Rota-Stabelli, Nicolas Lartillot, Gert Wörheide, and Davide Pisani. 2017. "Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals." *Current Biology* 27 (24). Elsevier: 3864–70.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of Phyml 3.0." *Systematic Biology* 59 (3). Oxford University Press: 307–21.

Hansen, Andrea, Sabine Hansmann, Tahir Samigullin, Andrey Antonov, and William Martin. 1999. "Gnetum and the Angiosperms: Molecular Evidence That Their Shared Morphological Characters Are Convergent, Rather Than Homologous." *Molecular Biology and Evolution* 16 (7). Oxford University Press: 1006–6.

Hedges, S Blair, Joel Dudley, and Sudhir Kumar. 2006. "TimeTree: A Public Knowledge-Base of Divergence Times Among Organisms." *Bioinformatics* 22 (23). Oxford University Press: 2971–2.

Huang, Huateng, Jeet Sukumaran, Stephen A Smith, and LLacey Knowles. 2017. "Cause of Gene Tree Discord? Distinguishing Incomplete Lineage Sorting and Lateral Gene Transfer in Phylogenetics." *PeerJ PrePrints* 5. PeerJ, Inc.: e3489v1.

Huelsenbeck, John P, JJ Bull, and Clifford W Cunningham. 1996. "Combining Data in Phylogenetic Analysis." *Trends in Ecology & Evolution* 11 (4). Elsevier: 152–58.

Jarvis, Erich D, Siavash Mirarab, Andre J Aberer, Bo Li, Peter Houde, Cai Li, Simon YW Ho, et al. 2014. "Whole-Genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds." *Science* 346 (6215). American Association for the Advancement of Science: 1320–31.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas KF Wong, Arndt von Haeseler, and Lars S Jermiin. 2017. "ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates." *Nature Methods* 14 (6). Nature Publishing Group: 587.

Karol, Kenneth G., Kathiravetpillai Arumuganathan, Jeffrey L. Boore, Aaron M. Duffy, Karin DE Everett, John D. Hall, S. Kellon Hansen, et al. 2010. "Complete Plastome Sequences of Equisetum Arvense and Isoetes Flaccida: Implications for Phylogeny and Plastid Genome Evolution of Early Land Plant Lineages." *BMC Evolutionary Biology* 10 (1): 321. https://doi.org/10.1186/1471-2148-10-321.

Kluge, Arnold G. 1989. "A Concern for Evidence and a Phylogenetic Hypothesis of Relationships Among Epicrates (Boidae, Serpentes)." *Systematic Biology* 38 (1). Society of Systematic Zoology: 7–25.

Knowles, L Lacey, Huateng Huang, Jeet Sukumaran, and Stephen A Smith. 2018. "A Matter of Phylogenetic Scale: Distinguishing Incomplete Lineage Sorting from Lateral Gene Transfer as the Cause of Gene Tree Discord in Recent Versus Deep Diversification Histories." *American Journal of Botany* 105 (3). Wiley Online Library: 376–84.

Kubatko, Laura Salter. 2009. "Identifying Hybridization Events in the Presence of Coalescence via Model Selection." *Systematic Biology* 58 (5). Oxford University Press: 478–88.

Lanfear, Robert, Paul B Frandsen, April M Wright, Tereza Senfeld, and Brett Calcott. 2016. "Partition-Finder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses." *Molecular Biology and Evolution* 34 (3). Oxford University Press: 772–73.

Liu, Liang, Lili Yu, Dennis K Pearl, and Scott V Edwards. 2009. "Estimating Species Phylogenies Using Coalescence Times Among Sequences." *Systematic Biology* 58 (5). Oxford University Press: 468–77.

Maddison, Wayne P. 1997. "Gene Trees in Species Trees." *Systematic Biology* 46 (3). Oxford University Press: 523–36.

Mirarab, Siavash, Md Shamsuzzoha Bayzid, Bastien Boussau, and Tandy Warnow. 2014. "Statistical Binning Enables an Accurate Coalescent-Based Estimation of the Avian Tree." *Science* 346 (6215). American Association for the Advancement of Science: 1250463.

Mirarab, Siavash, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. 2014. "ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation." *Bioinformatics* 30 (17). Oxford University Press: i541–i548.

Neupane, Suman, Karolina Fucikova, Louise A Lewis, Lynn Kuo, Ming-Hui Chen, and Paul Lewis. 2018. "Assessing Combinability of Phylogenomic Data Using Bayes Factors." *bioRxiv*. Cold Spring Harbor Laboratory, 250969.

Nguyen, Lam-Tung, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2014. "IQ-Tree: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32 (1). Oxford University Press: 268–74.

Nickrent, Daniel L., Christopher L. Parkinson, Jeffrey D. Palmer, and R. Joel Duff. 2000. "Multigene Phylogeny of Land Plants with Special Reference to Bryophytes and the Earliest Land Plants." *Molecular Biology and Evolution* 17 (12): 1885–95. https://doi.org/10.1093/oxfordjournals.molbev.a026290.

Nishiyama, T, and M Kato. 1999. "Molecular Phylogenetic Analysis Among Bryophytes and Tracheophytes Based on Combined Data of Plastid Coded Genes and the 18S rRNA Gene." *Molecular Biology and Evolution* 16 (8): 1027–36. https://doi.org/10.1093/oxfordjournals.molbev.a026192.

Penny, D, MD Hendy, EA Zimmer, and RK Hamby. 1990. "Trees from Sequences: Panacea or Pandora's Box?" *Australian Systematic Botany* 3 (1). CSIRO: 21–38.

Puttick, Mark N, Jennifer L Morris, Tom A Williams, Cymon J Cox, Dianne Edwards, Paul Kenrick, Silvia Pressel, et al. 2018. "The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte." *Current Biology* 28 (5). Elsevier: 733–45.

Qiu, Yin-Long, Libo Li, Bin Wang, Zhiduan Chen, Volker Knoop, Milena Groth-Malonek, Olena Dombrovska, et al. 2006. "The Deepest Divergences in Land Plants Inferred from Phylogenomic Evidence." *Proceedings of the National Academy of Sciences* 103 (42). National Academy of Sciences: 15511–6. https://doi.org/10.1073/pnas.0603335103.

Robinson, David F, and Leslie R Foulds. 1981. "Comparison of Phylogenetic Trees." *Mathematical Biosciences* 53 (1-2). Elsevier: 131–47.

Sauquet, Hervé, Maria von Balthazar, Susana Magallón, James A Doyle, Peter K Endress, Emily J Bailes, Erica Barroso de Morais, et al. 2017. "The Ancestral Flower of Angiosperms and Its Early Diversification." *Nature Communications* 8. Nature Publishing Group: 16047.

Sayou, Camille, Marie Monniaux, Max H Nanao, Edwige Moyroud, Samuel F Brockington, Emmanuel Thévenon, Hicham Chahtane, et al. 2014. "A Promiscuous Intermediate Underlies the Evolution of Leafy Dna Binding Specificity." *Science* 343 (6171). American Association for the Advancement of Science: 645–48.

Shen, Xing-Xing, Chris Todd Hittinger, and Antonis Rokas. 2017. "Contentious Relationships in Phylogenomic Studies Can Be Driven by a Handful of Genes." *Nature Ecology & Evolution* 1 (5). Nature Publishing Group: 0126.

Shen, Xing-Xing, Xiaofan Zhou, Jacek Kominek, Cletus P Kurtzman, Chris Todd Hittinger, and

Antonis Rokas. 2016. "Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data." *G3: Genes, Genomes, Genetics* 6 (12). G3: Genes, Genomes, Genetics: 3927–39.

Shimodaira, Hidetoshi, and Masami Hasegawa. 1999. "Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference." *Molecular Biology and Evolution* 16 (8). Oxford University Press: 1114–4.

Simion, Paul, Hervé Philippe, Denis Baurain, Muriel Jager, Daniel J. Richter, Arnaud Di Franco, Béatrice Roure, et al. 2017. "A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals." *Current Biology* 27 (7): 958–67. https://doi.org/https://doi.org/10.1016/j.cub.2017.02.031.

Smith, Stephen A, Joseph W Brown, and Joseph F Walker. 2018. "So Many Genes, so Little Time: A Practical Approach to Divergence-Time Estimation in the Genomic Era." *PloS One* 13 (5). Public Library of Science: e0197433.

Smith, Stephen A, Michael J Moore, Joseph W Brown, and Ya Yang. 2015. "Analysis of Phylogenomic Datasets Reveals Conflict, Concordance, and Gene Duplications with Examples from Animals and Plants." *BMC Evolutionary Biology* 15 (1). BioMed Central: 150.

Springer, Mark S, and John Gatesy. 2016. "The Gene Tree Delusion." *Molecular Phylogenetics and Evolution* 94. Elsevier: 1–33.

Theobald, Douglas L. 2010. "A Formal Test of the Theory of Universal Common Ancestry." *Nature* 465 (7295). Nature Publishing Group: 219.

Villarreal, Juan Carlos, and Susanne S Renner. 2012. "Hornwort Pyrenoids, Carbon-Concentrating Structures, Evolved and Were Lost at Least Five Times During the Last 100 Million Years." *Proceedings of the National Academy of Sciences* 109 (46). National Acad Sciences: 18873–8.

Walker, Joseph F, Joseph W Brown, and Stephen A Smith. 2018. "Analyzing Contentious Relationships and Outlier Genes in Phylogenomics." *Systematic Biology* syy043. Oxford University Press. https://doi.org/10.1093/sysbio/syy043.

Walker, Joseph F, Ya Yang, Tao Feng, Alfonso Timoneda, Jessica Mikenas, Vera Hutchison, Caroline Edwards, et al. 2018. "From Cacti to Carnivores: Improved Phylotranscriptomic Sampling and Hierarchical Homology Inference Provide Further Insight into the Evolution of Caryophyllales." *American Journal of Botany* 105 (3). Wiley Online Library: 446–62.

Whelan, Nathan V, Kevin M Kocot, Tatiana P Moroz, Krishanu Mukherjee, Peter Williams, Gustav Paulay, Leonid L Moroz, and Kenneth M Halanych. 2017. "Ctenophore Relationships and Their Placement as the Sister Group to All Other Animals." *Nature Ecology & Evolution* 1 (11). Nature Publishing Group: 1737.

Wickett, Norman J, Siavash Mirarab, Nam Nguyen, Tandy Warnow, Eric Carpenter, Naim Matasci, Saravanaraj Ayyampalayam, et al. 2014. "Phylotranscriptomic Analysis of the Origin and Early Diversification of Land Plants." *Proceedings of the National Academy of Sciences* 111 (45). National Acad Sciences: E4859–E4868.

Xi, Zhenxiang, Liang Liu, Joshua S Rest, and Charles C Davis. 2014. "Coalescent Versus Concatenation Methods and the Placement of Amborella as Sister to Water Lilies." *Systematic Biology* 63 (6). Oxford University Press: 919–32.

Yang, Ya, and Stephen A Smith. 2014. "Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics." *Molecular Biology and Evolution* 31 (11). Oxford University Press: 3081–92.