



# Characterizing gene tree conflict in plastome-inferred phylogenies

Joseph F. Walker<sup>1,\*</sup>, Nathanael Walker-Hale<sup>2,\*</sup>, Oscar M. Vargas<sup>3</sup>, Drew A. Larson<sup>4</sup> and Gregory W. Stull<sup>5</sup>

<sup>1</sup>Sainsbury Laboratory (SLCU), University of Cambridge, Cambridge, United Kingdom

<sup>2</sup>Department of Plant Sciences, University of Cambridge, Cambridge, Cambridgeshire, United Kingdom

<sup>3</sup>University of California, Santa Cruz, Santa Cruz, United States of America

<sup>4</sup>University of Michigan—Ann Arbor, Ann Arbor, MI, United States of America

<sup>5</sup>Department of Botany, Smithsonian Institution, Washington, United States of America

\*These authors contributed equally to this work.

## ABSTRACT

Evolutionary relationships among plants have been inferred primarily using chloroplast data. To date, no study has comprehensively examined the plastome for gene tree conflict. Using a broad sampling of angiosperm plastomes, we characterize gene tree conflict among plastid genes at various time scales and explore correlates to conflict (e.g., evolutionary rate, gene length, molecule type). We uncover notable gene tree conflict against a backdrop of largely uninformative genes. We find alignment length and tree length are strong predictors of concordance, and that nucleotides outperform amino acids. Of the most commonly used markers, *matK*, greatly outperforms *rbcL*; however, the rarely used gene *rpoC2* is the top-performing gene in every analysis. We find that *rpoC2* reconstructs angiosperm phylogeny as well as the entire concatenated set of protein-coding chloroplast genes. Our results suggest that longer genes are superior for phylogeny reconstruction. The alleviation of some conflict through the use of nucleotides suggests that stochastic and systematic error is likely the root of most of the observed conflict, but further research on biological conflict within plastome is warranted given documented cases of heteroplasmic recombination. We suggest that researchers should filter genes for topological concordance when performing downstream comparative analyses on phylogenetic data, even when using chloroplast genomes.

Submitted 20 May 2019  
Accepted 25 August 2019  
Published 24 September 2019

Corresponding authors  
Joseph F. Walker, jfw52@cam.ac.uk  
Gregory W. Stull, gwstull@gmail.com

Academic editor  
Scott Edwards

Additional Information and  
Declarations can be found on  
page 21

DOI 10.7717/peerj.7747

© Copyright  
2019 Walker et al.

Distributed under  
Creative Commons CC-BY 4.0

OPEN ACCESS

**Subjects** Evolutionary Studies, Plant Science

**Keywords** Angiosperms, Chloroplast, Gene tree conflict, *rpoC2*, Phylogenomics, Plastome, *matK*, *rbcL*

## INTRODUCTION

Chloroplast data have been the most prominent source of information for plant phylogenetics, largely due to the ease with which chloroplast genes can be sequenced, assembled, and analyzed (Palmer, 1985; Taberlet et al., 1991). The majority of broad-scale phylogenetic studies on plants have used chloroplast genes (e.g., Chase et al., 1993; Soltis et al., 2000; Soltis et al., 2011), and the resulting phylogenies have been used for countless comparative studies examining ancestral states, historical biogeography, and other evolutionary patterns. While older studies relied mostly on targeted genes such as

*rbcl* and *matK*, recent advances in DNA sequencing have drastically increased the ease and affordability of whole-chloroplast genome (i.e., plastome) sequencing (Moore et al., 2006; Cronn et al., 2008; Cronn et al., 2012; Stull et al., 2013; Uribe-Convers et al., 2014), increasing the number of studies employing plastome-scale data for phylogenetic and comparative analyses (e.g., Jansen et al., 2007; Moore et al., 2007; Moore et al., 2010; Ruhfel et al., 2014; Stull et al., 2015; Gitzendanner et al., 2018; Li et al., 2019). Nonetheless, the utility of plastid genes, as well as the entire plastome, is ultimately determined by the extent to which they reflect ‘true’ evolutionary relationships (i.e., the ‘species tree’) of the lineages in question (Doyle, 1992).

Phylogenomic conflict (i.e., the presence of conflicting relationships among gene trees in a genomic dataset) is now recognized as a nearly ubiquitous feature of nuclear phylogenomic studies (Smith et al., 2015; Rokas et al., 2003). Gene tree conflict is commonly attributed to biological causes such as incomplete lineage sorting, hybridization, and gene duplication and loss (Maddison, 1997; Galtier & Daubin, 2008; Smith et al., 2015; Walker et al., 2017; Vargas, Ortiz & Simpson, 2017). The genes within the plastome, however, are generally thought to be free of such biological sources of conflict. This is because the plastome is typically uniparentally inherited (maternally in angiosperms, and at times paternally in conifers: Mogensen, 1996; Birky Jr, 1995) and undergoes a unique form of recombination that is not expected to result in conflicting gene histories within the plastome (Palmer, 1983; Bendich, 2004; Walker et al., 2015).

In angiosperms, nonmaternal inheritance of the chloroplast has been reported (e.g., Smith, 1989; McCauley et al., 2007), and several surveys of pollen in flowering plants (Corriveau & Coleman, 1988; Zhang, Liu & Sodmergen, 2003) have documented plastid DNA in up to 18% of the species examined, indicating potential for biparental inheritance and heteroplasmy (i.e., the presence of two or more different plastomes in a single organism, cell, or organelle). Heteroplasmy, which has been documented in multiple angiosperm species (e.g., Johnson & Palmer, 1988; Lee, Blake & Smith, 1988; Hansen et al., 2007; Carbonell-Caballero et al., 2015), creates an opportunity for heteroplasmic recombination, which could result in gene tree conflict in plastome-inferred phylogenies. Unlike animal organelles, where there is little evidence for heteroplasmic recombination (White et al., 2008), and thus a strong justification for analyzing the mitochondrial genome as a single unit (Richards et al., 2018), heteroplasmic recombination (both intra- and interspecific) has been invoked by multiple studies to explain discordance in various plant clades (Huang et al., 2001; Marshall, Newton & Ritland, 2001; Erixson & Oxelman, 2008; Bouillé, Senneville & Bousquet, 2011). However, only a few recent studies have documented clear evidence of this phenomenon (Sullivan et al., 2017; Sancho et al., 2018). Beyond heteroplasmy, sharing of genes between the chloroplast and nuclear or mitochondrial genomes remains another potential source of biological conflict (Martin et al., 1998; Martin, 2003; Stegemann et al., 2003; Rice et al., 2013; Straub et al., 2013; Smith, 2014), although transfer back to the plastome would be required to generate conflict and this has rarely been found (Smith, 2014). Even though biological conflict in the plastome generally seems rare, the full extent of intraplasmic

conflict has yet to be evaluated. Quantifying its extent is of high importance given that most studies assume no conflict and treat the plastome as a single locus (Wolfe & Randle, 2004).

Aside from biological sources of conflict, there also remain significant potential sources of systematic and stochastic error that could contribute to conflict across the plastome (e.g., Burleigh & Mathews, 2007a; Burleigh & Mathews, 2007b). Chloroplast data are used at various time scales, and the accumulation of substitutions over long periods of evolutionary time increases the probability of encountering systematic error due to saturation (Rodríguez-Ezpeleta et al., 2007; Philippe et al., 2011). Conflict has been demonstrated among different functional groups of genes (Liu et al., 2012), among different regions of the plastome (Davis et al., 2013; Walker, Zanis & Emery, 2014), as well as among individual genes (e.g., Shepherd, Holland & Perrie, 2008; Foster, Henwood & Ho, 2018; Gonçalves et al., 2019). The rate of chloroplast evolution as a whole has been examined (and compared with the nuclear and mitochondrial genomes; Wolfe, Li & Sharp, 1987), and rate variation within the chloroplast—especially across the three major regions of the genome, i.e., the large single-copy (LSC) region, the small single-copy (SSC) region, and the inverted repeats (IRa, IRb)—has been explored to help determine the markers useful for phylogenetic inference at different time scales (e.g., Graham & Olmstead, 2000; Shaw et al., 2005; Shaw et al., 2007; Shaw et al., 2014). However, no study has comprehensively examined gene tree conflict within the plastome to better characterize the extent and sources of conflict, and to identify the plastid genes most concordant with our current understanding of angiosperm phylogeny (e.g., Soltis et al., 2011; Wickett et al., 2014; Zeng et al., 2014; Gitzendanner et al., 2018; but see Logacheva et al., 2007 for a preliminary investigation of the concordance of individual plastid genes with angiosperm phylogeny).

Here we use phylogenomic tools to characterize the extent of conflict among plastid genes as a function of evolutionary rate, rate variation among species, sequence length, and data type (i.e., nucleotides vs. amino acids) at varying time scales across angiosperms. Our results show that plastid genes, when analyzed in combination, agree with many community consensus relationships; however, across plastid genes there is notable gene tree conflict. The number of conflicting genes at each node is often comparable to the number of concordant genes; however, the majority of plastid genes are uninformative for most nodes when considering support and thus unlikely to positively mislead studies. Information content (gene length and molecule type, i.e., nucleotides vs amino acids) and evolutionary rate were positive predictors of concordance, suggesting that most observed gene-tree conflict is a consequence of spurious inferences from insufficient information. However, some nodes across angiosperm phylogeny show numerous strongly supported conflicting genes; this rare event indicates an area for future exploration. This may result from systematic error or could be the product of a biological source of gene-tree conflict (e.g., heteroplasmic recombination, horizontal gene transfer). We also document the performance of individual genes at recapitulating angiosperm phylogeny, finding the seldom-used gene *rpoC2* to outperform commonly used genes (e.g., *rbcL*, *matK*) in all cases, consistent with previous work highlighting the utility of this gene (Logacheva et al., 2007). Although as a whole the chloroplast infers a topology similar to currently hypothesized angiosperm relationships, the conflict underlying it can influence branch

lengths, and therefore dating analyses, comparative methods, and other work that relies on branch lengths could benefit from gene filtering procedures. Our results provide insight into the extent and sources of intraplastome conflict.

## MATERIALS AND METHODS

### Data acquisition and sampling

Complete plastome coding data (both nucleotide and amino acid) were downloaded from NCBI for 53 taxa: 51 angiosperm ingroups and two gymnosperm outgroups (*Ginkgo biloba* and *Podocarpus lambertii*; Table S1). Our sampling scheme was designed to capture all major angiosperm lineages (e.g., *Soltis et al., 2011*) thus allowing accurate inference of angiosperm phylogeny, while also including denser sampling for nested clades in Asterales. This allowed us to evaluate the extent of gene tree conflict at different evolutionary levels/time scales, from species-level relationships in *Diplostephium* (Asteraceae) to the ordinal-level relationships defining the backbone of angiosperm phylogeny.

### Data preparation, alignment, and phylogenetic inference

All scripts and data used may be found on GitHub (<https://github.com/jfwalker/ChloroplastPhylogenomics>). Orthology was determined based upon the annotations of protein-coding genes on Genbank; this resulted in near-complete gene occupancy apart from instances of gene loss or reported pseudogenization. We did not use non-coding regions, as proper orthology at deep time scales can be difficult to assess due to gene rearrangements and inversions; this is in part why most deep phylogenetic analyses of angiosperms using plastomes have excluded non-coding data (e.g., *Moore et al., 2010*; *Ruhfel et al., 2014*).

The amino acid and nucleotide data were aligned using Fast Statistical Alignment (FSA v1.15.9; *Bradley et al., 2009*) with the default settings for peptide and the setting “-noanchored” for nucleotide. FSA has been shown to be one of the top-performing alignment programs (*Bradley et al., 2009*; *Redelings, 2014*), and does not rely upon a guide tree for sequence alignment, alleviating downstream bias (*Bradley et al., 2009*; *Boyce, Sievers & Higgins, 2015*; *Chatzou et al., 2018*). A maximum likelihood (ML) tree was then inferred for each gene using RAxMLv.8.2.4 (*Stamatakis, 2014*), with the PROTGAMMAAUTO and GTR+G models of evolution used for the amino acid and nucleotide data, respectively. For each dataset, we conducted 200 rapid bootstrap replicates. The alignments were also concatenated into supermatrices and partitioned by gene using the phyx (v0.99) program pxcats (*Brown, Walker & Smith, 2017*).

The nucleotide and amino acid supermatrices were then each used to infer concatenated ‘plastome’ trees using the GTR+G and the PROTGAMMAAUTO models, respectively, as implemented in RAxML. To complement the model inference performed by RAxML from the AUTO feature, we also used IQ-TREE’s (v1.5.4) (*Nguyen et al., 2014*) built-in model selection process (*Kalyaanamoorthy et al., 2017*) on the partitioned data. We did not partition by codon position because the lengths of most of the plastid genes, especially when divided into three smaller partitions, would have been insufficient to inform an evolutionary model.

Recent work has advocated the use of multispecies coalescent methods when analyzing sets of gene trees from the chloroplast (Gonçalves *et al.*, 2019). As such we reanalyzed the nucleotide data using the Maximum Quartet Support Species Tree method as implemented in ASTRAL v5.6.3 (Zhang *et al.*, 2018) with default settings.

### Reference phylogenies

For the conflict analyses, described below, we created several reference trees against which the gene trees were mapped. Primarily, we used a topology based on Soltis *et al.* (2011) and where possible that of Wickett *et al.* (2014), and—for species-level relationships within Asteraceae—Vargas, Ortiz & Simpson (2017). We refer to this as our ‘accepted tree’, or AT (Fig. 1). We also created a reference tree based on the recent plastome phylogeny of Gitzendanner *et al.* (2018), which we refer to as the ‘Gitzendanner tree’, or GT. The GT is highly concordant with the AT, but several deep (and commonly contentious) nodes differ, e.g., the placement of Buxales and ordinal relationships in lamiids (here, we only sampled Lamiales, Gentianales, and Solanales). In the latter case, all three possible combinations of lamiids taxa were assessed (discussed further below in Assessment of Supported Conflict). Although it is difficult to determine which of these trees better represents species relationships of angiosperms, they both serve as useful frameworks for revealing conflict among genes. Here, we primarily focus on analyses using the AT, but conflict analyses using the GT were also conducted to ensure that our overall results were consistent across reference trees.

### Analysis of conflict

All gene trees were rooted on the outgroups using the phyx program pxrr (Brown, Walker & Smith, 2017), in a ranked fashion (“-r”) in the order *Podocarpus lambertii* then *Ginkgo biloba*; this was performed to account for any missing genes in either of the taxa. In all cases the trees were rooted with at least one gymnosperm. Gene tree conflict in the data was identified using the bipartition method as implemented in phyparts v.0.0.1 (Smith *et al.*, 2015), with the gene trees from each data set (amino acid and nucleotide) mapped against the AT and GT. The concordance analyses were performed using both a support cutoff (at  $\geq 70\%$  bootstrap support, i.e., moderate support as suggested by Hillis & Bull, 1993) and no support cutoff. When the support cutoff is used, any gene tree node with less than 70% bootstrap support is regarded as uninformative for the reference node in question (i.e., it is uncertain whether the gene tree node is in conflict or concordance with the reference node); when no support cutoff is used, the gene tree node is evaluated as conflicting or concordant regardless of the support value. However, in both cases, a gene is considered uninformative if a taxon relevant to a particular node/relationship is missing from the gene dataset. For example, consider the reference tree ((A,B),C); if a gene tree is missing taxon B, but contains the relationship (A,C), then the gene tree is concordant with the node ((A,B),C) but uninformative regarding the node (A,B). See Smith *et al.* (2015) for an in-depth explanation of this method. To examine patterns of concordance and conflict among all gene trees, we calculated the pairwise Robinson-Foulds distance (Robinson & Foulds, 1981) (RF) between each gene tree and the AT, concatenated topology,

and ASTRAL tree. We created a network of trees based on RF distance using the R package *igraph* v1.2.4.1 in R v3.6.1 (Csardi & Nepusz, 2006; R Core Team, 2018), using trees as nodes, inverse RF distances as edge weights, and the Fruchterman-Reingold algorithm to lay out the graph (Fruchterman & Reingold, 1991).

Levels of conflict were examined across the entire tree and within particular time intervals of the evolutionary history of angiosperms. For the latter case, the ~150 Ma of crown angiosperm evolution were divided into five time-intervals (30 Ma each): 150 to 121 mya, 120 to 91 mya, 90 to 61 mya, 60 to 31 mya, and 30 to 0 mya. The nodes across the tree were binned into the time intervals based on their inferred ages (ages >20 Ma from Magallón *et al.*, 2015; ages <20 Ma from Vargas, Ortiz & Simpson, 2017; Roquet *et al.*, 2009). At each time interval, the proportion of concordant nodes for each gene was calculated (for all the nodes falling within that time interval). This allowed us to assess the level(s) of divergence at which each gene is most informative. We also determined concordance levels of the ‘plastome’ trees (against the reference) within each of these time intervals.

### Assessments of individual plastid genes

Using the *phyx* program *pxlstr* (Brown, Walker & Smith, 2017), we calculated summary statistics for each gene alignment and corresponding gene tree: number of included species (as gene alignments could be missing a species from pseudogenization or gene loss), alignment length, tree length (a measure of gene evolutionary rate), and root-to-tip variance (a measure of rate variation across the phylogeny). Alignment length and tree length represent complementary measures of a gene’s information content. All else being equal, longer genes should harbor more phylogenetic signal. Tree length serves as a model-informed metric because all sites, and not just those which are variable or parsimony informative, are used in estimation of model distances. Because our study uses maximum likelihood methods, we chose this model-informed metric as opposed to aforementioned metrics. Levels of concordance of each gene tree with the reference trees were then assessed by tabulating the number of nodes concordant between the gene tree and the given reference tree (e.g., the AT). The number of concordant nodes (in Fig. 2 treated as a proportion of total nodes available to support and in Fig. 3 based on total nodes) was used as a measure of a gene’s ability to accurately reconstruct the angiosperm phylogeny.

### Predictors of concordance

We examined the statistical relationships between gene tree concordance (using concordance data from the AT-based conflict analysis) and alignment length, tree length, and root-to-tip variance. We treated each node as a trial and analyzed the concordance (success) or conflict (failure) in aggregate, such that each gene was a single observation with  $x$  successes in  $n$  trials. Because of variable sampling and loss of some genes,  $n$  was variable between observations. We analyzed the relationship between probability of concordance and alignment length, tree length, and root-to-tip variance using logistic regression of aggregate binomial trials with the function *glm()* in R v3.4.4

(*R Core Team, 2018*). Binomial models were generally characterized by high residual deviance, and we thus allowed for overdispersion by fitting quasibinomial logistic regressions (using ‘family = quasibinomial()’ in R v3.4.4). We fitted models for both amino acid and nucleotide data, both not considering and considering support, and for both alignment methods. We also fitted models for amino acid data inferred only under WAG.

We modeled gene performance as a function of length, tree length, and root-to-tip variation, and as a function of each predictor individually. Because it is possible that apparent relationships between alignment length and concordance may reflect signal from gene information content per alignment site, we also modeled gene performance as a function of length and tree length (as a proxy of gene information content and rate, respectively, as noted above), to assess the relationship between alignment length and gene performance after controlling for variation associated with rate.

Investigation of model fits on full datasets by studentized residuals, leverage, and Cook’s distance values revealed that several observations were highly influential. Therefore, we also conducted analyses on reduced datasets to investigate the influence of these observations. In the amino acid datasets, we excluded *rpl22* and *rpl32*, which were probably influential due to their high tree length values, and *ycf1* and *ycf2*, which were probably influential based on their long alignments. In the nucleotide datasets, we excluded *clpP*, which had poor performance relative to tree and alignment length, *rps15*, which had a relatively long tree length, and *ycf1* and *ycf2* for the same reasons as previously noted. Combined analyses of alignment length and tree length were not subject to influence driven by high root-to-tip variance values, and hence reduced datasets had fewer genes removed. In this case, we excluded only *ycf1* and *ycf2*. Regression results were summarized in tables using the R package Stargazer (*Hlavac, 2018*).

### Saturation analyses

We also performed saturation analyses on all the chloroplast genes to determine if they were capable of inferring deep divergences (*Philippe et al., 1994*). In this analysis, pairwise distances between sequences are compared to model-corrected distances as a means of identifying if the real genetic distances may be properly estimated. A saturation analysis was performed on each gene and each codon position and thus the data were realigned by codons for this analysis. The amino acid alignments were used to guide the codon alignments using the program pxaa2cdn, from the phyx package. Four of the genes did not match their corresponding amino acid sequence in its entirety and these were left out of this analysis, since the codons could not be properly aligned. This discrepancy is likely due to errors in the GenBank submission (the amino acid and nucleotide data do not perfectly correspond). Saturation was assessed by determining the observed number of differences between sequences compared to the inferred number of substitutions. This analysis was performed using the “dist.dna” and “dist.corrected” functions in the R package “ape” (*Paradis & Schliep, 2018*), with the F84 model of evolution (*Felsenstein & Churchill, 1996*) used for the correction.

## Comparison of genomic regions

We assessed the utility of the three major plastome regions—the Large Single Copy (LSC) region, Small Single Copy (SSC) region, and the Inverted Repeat (IR) region—for reconstructing angiosperm phylogeny in two ways. First, we constructed ML phylogenies by concatenating all genes found in each region (as described above for each of the plastome trees inferred using all genes). We then calculated the number of concordant nodes with the AT for each genomic region. Second, using the concordance levels of each individual gene (described above), we created a plastome diagram (with genes arranged according to their genomic position) showing the concordance levels of each gene at the five different time scales discussed above (Fig. S1); this permits a qualitative visual assessment of the general concordance levels of each genomic region at each time slice.

## Assessment of supported conflict

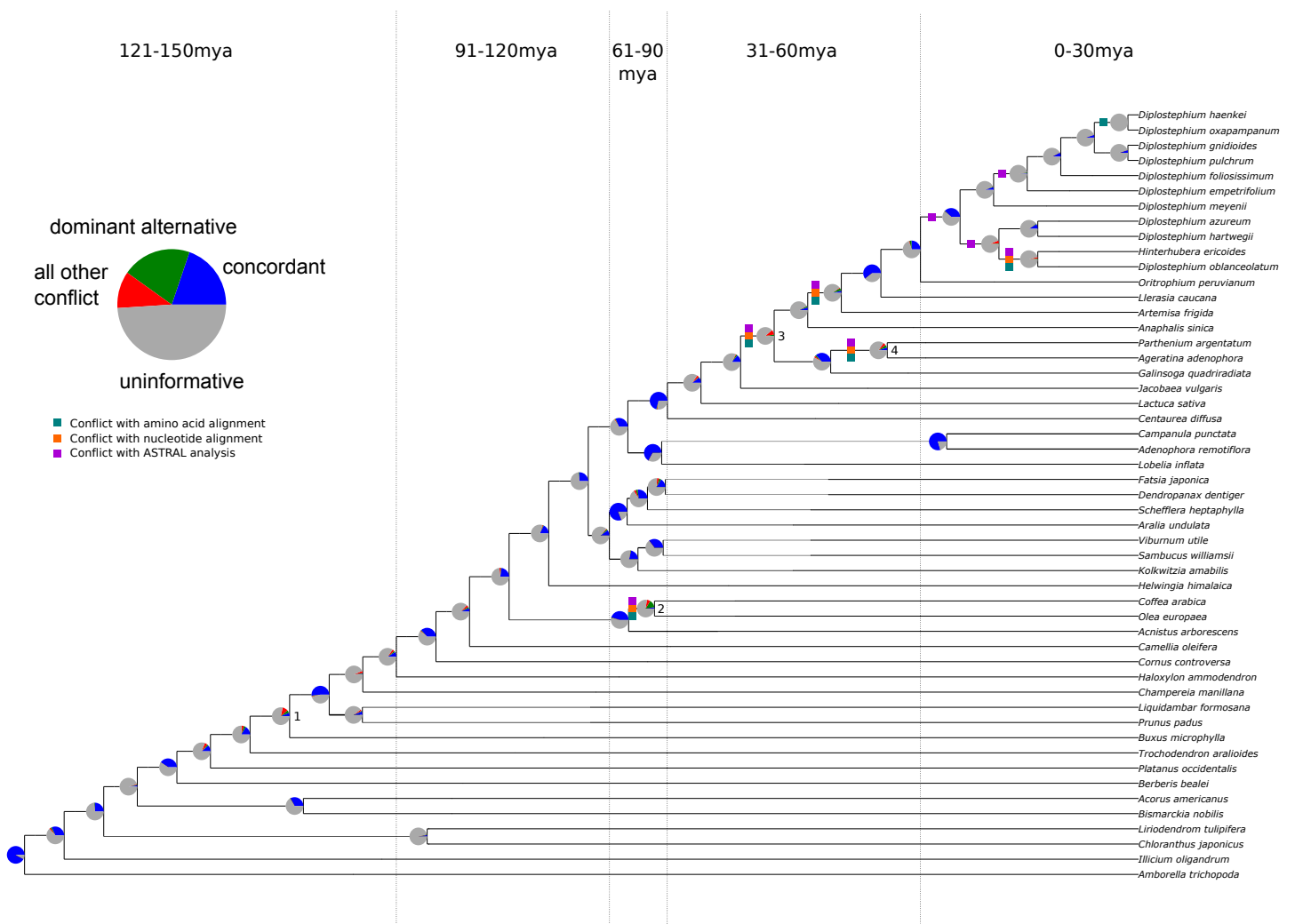
We compared four contentious regions of the AT to a series of alternative relationships. The tested alternative relationships were present in either the GT (i.e., the plastome phylogeny of *Gitzendanner et al., 2018*), our plastome phylogenies (with  $\geq 70$  BS for the node in question; we refer to this tree below as the ‘plastome tree’, or PT), and/or three or more of the gene trees from our nucleotide dataset (with  $\geq 70$  BS for the node in question). These alternative relationships pertain to: (1) the placements of *Buxus* and *Trochodendron*; (2) relationships among *Acnistus* (Solanales), *Coffea* (Gentianales), and *Olea* (Lamiales); (3) the placements of *Jacobaea* and *Artemisia* in Asteraceae; and (4) the placements of *Galinsoga*, *Ageratina*, and *Parthenium* in Asteraceae. To test these alternative relationships, we used a modified version of the Maximum Gene Wise Edge (MGWE) method (*Walker, Brown & Smith, 2018*), where instead of a defined “TREE SET”, we used a constraint tree for each alternative hypothesis. The likelihood for every gene was calculated across each constraint, thereby forcing the relationship in question to be the same. Similar to MGWE, this allowed for conflict across the rest of the topology; however, this provides a greater amount of tree space to be explored (similar to *Smith et al., 2018*). This method creates likelihood scores solely based on each alternative relationship, making it robust to gene tree conflict outside of the node in question.

The modification of the MGWE method has been implemented in the EdgeTest.py program of the package PHylogenetic Analysis Into Lineages (<https://github.com/jfwalker/PHAIL>).

## Assessment of influence of alignment method

Because initial homology statements at the alignment stage can strongly influence the outcome of phylogenetic inference (*Morrison, 2009; Morrison, Morgan & Kelchner, 2015*), we repeated the conflict analysis using the same parameters as above with two alternate sets of alignments. In one case we re-inferred using MAFFT v2.271 (*Katoh & Standley, 2013*) with the setting “-auto -max\_iterate 1,000”. In the second case we cleaned the original FSA gene alignments using pxclsq (*Brown, Walker & Smith, 2017*) for a minimum of 30% column occupancy in order to remove gap-heavy and ambiguously aligned sites, and RAxML v8 was used.



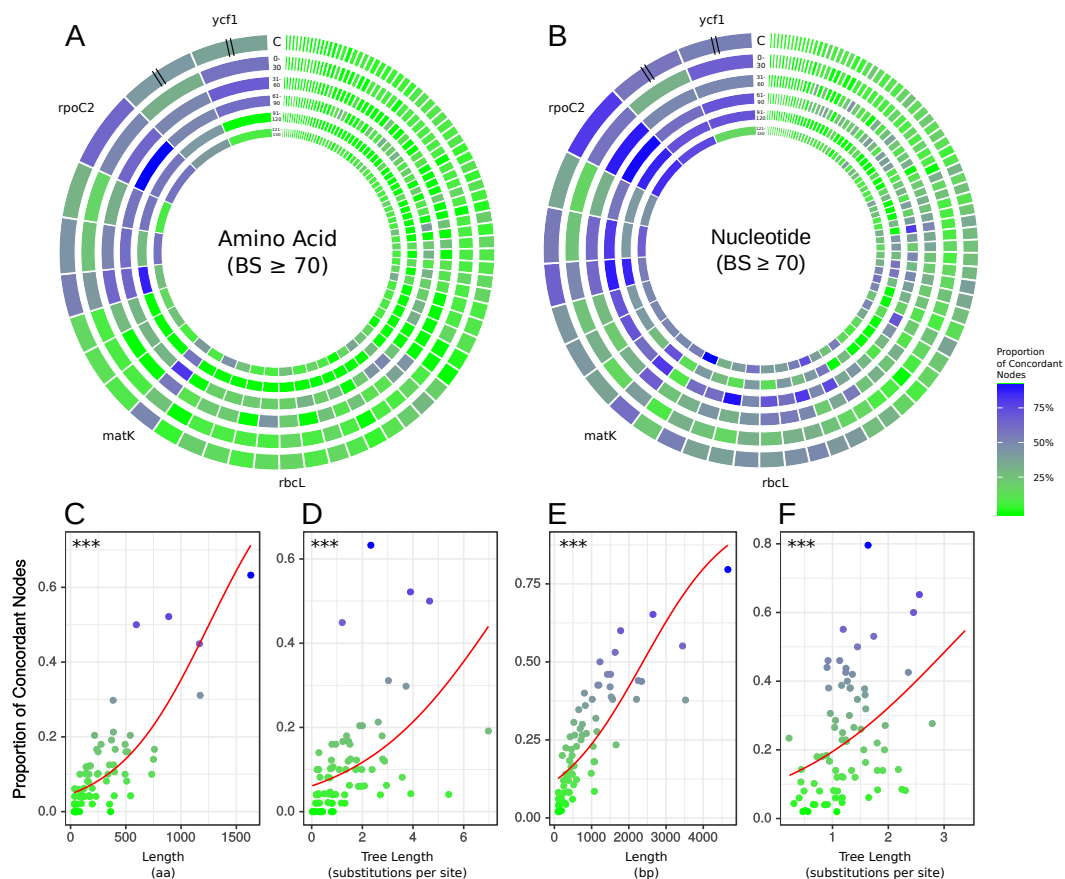


**Figure 1** Summary of chloroplast conflict against the reference phylogeny of angiosperms. Green, orange, and purple boxes indicate where the amino acid, nucleotide, or MQSST phylogenies conflict with the reference phylogeny. Pie charts depict the amount of gene tree conflict observed in the nucleotide analysis, with the blue, green, red and gray slices representing, respectively, the proportion of gene trees concordant, conflicting (supporting a single main alternative topology), conflicting (supporting various alternative topologies), and uninformative (BS < 70 or missing taxon) at each node in the species tree. The dashed lines represent 30 myr time intervals (positioned based on [Magallón et al., 2015](#) and [Vargas, Ortiz & Simpson, 2017](#)) used to bin nodes for examinations of conflict at different levels of divergence.

Full-size DOI: [10.7717/peerj.7747/fig-1](https://doi.org/10.7717/peerj.7747/fig-1)

## Assessment of influence of model

To test if the inferred model of evolution influenced the inference, we re-ran the tree inference using the same methods above on the uncleaned FSA amino acid datasets using only the WAG model of evolution ([Whelan & Goldman, 2001](#)). The WAG model was not found to be the optimal amino acid model for any gene in the analysis, thus this allowed us to assess the influence of assigning the same sub-optimal model across all partitions.



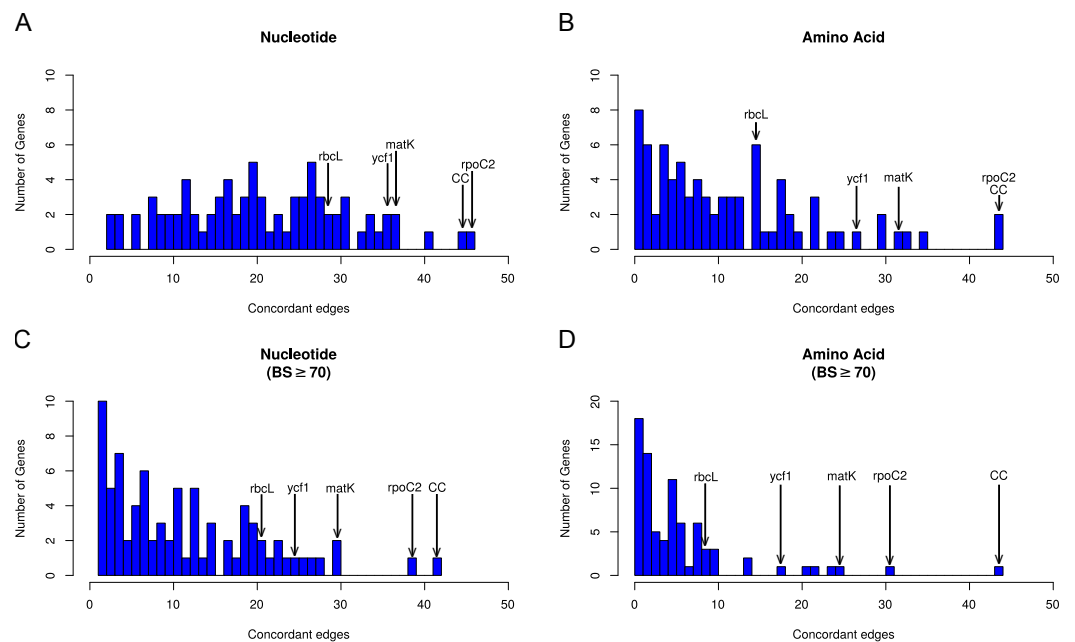
**Figure 2** Gene tree concordance/conflict at varying time scales. Each diagram represents a different molecule type and shows the proportion of concordance each gene exhibits at the five time slices shown in Fig. 1: (1) 150–121 mya, (2) 120–91 mya, (3) 90–61 mya, (4) 60–31 mya, (5) 30–0 mya and C is the concordance summed over all the time scales. The individual genes are scaled by length of alignment; however, *ycf1* and *ycf2* are cut to approximately the length of *rpoC2* due to their extremely long alignments. (A) Results from amino acid data considering only nodes with bootstrap support  $\geq 70\%$ . (B) Results from nucleotide data considering only nodes with bootstrap support  $\geq 70\%$ . The plots along the bottom show relationships between gene concordance levels and alignment length and tree length, excluding outlying genes (see Methods). Each point represents the proportion of concordance considering only nodes with bootstrap support  $\geq 70\%$ . Red lines show the predicted values from logistic regression and asterisks give the  $p$ -value of the relationship from univariate logistic regression, \*\*\* =  $p < 0.001$ . (C) Logistic regression of concordant nodes from amino acid data against alignment length. (D) Logistic regression of concordant nodes from amino acid data against tree length. (E) Logistic regression of concordant nodes from nucleotide data against alignment length. (F) Logistic regression of concordant nodes from nucleotide data against tree length.

Full-size DOI: [10.7717/peerj.7747/fig-2](https://doi.org/10.7717/peerj.7747/fig-2)

## RESULTS

### Influence of alignment and model

The use of MAFFT as opposed to FSA as an alignment tool did not produce a qualitatively different result (Tables S2–S7). Furthermore, our results were largely not qualitatively influenced by using an alignment that had been cleaned for 30% minimum occupancy (Tables S2–S6 and Tables S8–S9). Quasibinomial regression results for amino acids



**Figure 3** Histograms depicting number of concordant edges each gene tree contains compared to the reference phylogeny (i.e., the AT). Histograms are binned by number of concordant nodes; bar heights give the number of genes in each bin. Commonly used markers (*matK*, *ndhF*, *rbcL*, and *ycf1*) are labeled on the graph, along with the most concordant gene (*rpoC2*) and the number of concordant nodes for the complete chloroplast (CC) compared to the AT. (A) Concordance in the nucleotide dataset not considering bootstrap support. (B) Concordance in the amino acid dataset not considering bootstrap support. (C) Concordance in the nucleotide dataset considering bootstrap support  $\geq 70\%$ . (D) Concordance in the amino acid dataset considering bootstrap support  $\geq 70\%$ .

Full-size [DOI: 10.7717/peerj.7747/fig-3](https://doi.org/10.7717/peerj.7747/fig-3)

aligned with FSA and cleaned for 30% minimum occupancy were not significant for all predictors, but this model had an abnormally high estimated dispersion parameter (almost 17 times greater than the next largest estimate), suggesting that  $p$ -values inferred under this model were unreliable (Tables S2–S5). When running a sub-optimal model of evolution for all genes (the WAG model) we recapitulated the results of our model-tested analysis (Tables S2–S5 and Table S10). The differences among all these analyses were occasionally off by one in terms of number of concordant nodes; however, no results qualitatively changed: ranking of genes in terms of concordance, inference from regression models, and detection of strongly-supported conflict were largely unchanged. Therefore, for the remainder of the results and discussion we reference the uncleaned FSA alignment (Tables S11–S12).

### Patterns of conflicting chloroplast signal

The sampling for this experiment allowed us to examine conflict at multiple timescales (the values for each analysis may be found in Tables S6–S12), with roughly 10 divergences occurring in each 30 Ma bin. Our ‘plastome’ trees, inferred from the concatenated gene sets (nucleotide and amino acid), were highly concordant with the AT (Fig. 1 and Fig. S2). Without considering support, the gene trees showed high levels of conflict across the different analyses. This pattern is similar to that seen in nuclear phylogenomic

datasets (Salichos & Rokas, 2013; Smith et al., 2015; Shen, Salichos & Rokas, 2016). When considering support, the majority of the plastid genes were uninformative for practically all nodes in the phylogeny (Fig. 1; Tables S11–S12); i.e., they had bootstrap support below 70% (moderate support) for that particular relationship (whether in conflict or concordance).

There was no obvious relationship between the amount of gene tree conflict and evolutionary scale (i.e., conflict was relatively evenly distributed across shallow and deeper nodes/time scales; Figs. 1 and 2). Although the greatest degree of gene tree concordance with the AT appeared in the nodes with inferred ages between 90–61 Ma (ages based on Magallón et al., 2015), these nodes typically still contained at least 50% uninformative gene trees (Fig. 1). Genome location also showed weak correspondence to conflict/concordance (Fig. S1; Table S13); discussed further below. Rather than timescale and genome location, data type (nucleotide vs amino acid) had a much greater impact on the prevalence of conflict (Figs. 2 and 3), with the amino acid dataset generally showing higher levels of gene tree conflict. When factoring in support (BS  $\geq$  70% cutoff), the amino acid data set showed even less concordance with the AT (as more genes were considered uninformative due to low BS support). Considering support also decreased concordance of the nucleotide data set with the AT, but proportionally less. See Figs. S2–S3 for conflict analyses showing the amino acid and nucleotide gene trees mapped onto the AT (the nucleotide results from Fig. S2 are also shown in Fig. 1).

To identify how many different models of amino acid evolution underlie the genes in the plastome, we tested each gene against the candidate set of amino acid models in IQ-TREE and RAxML. We found that a wide range of evolutionary models best fit the individual genes—rather than just a single model that best fits every gene (Table S12). Many of the models were not designed specifically for plastome data, and cpRev, which was designed for plastome data, was only the best fit for 19 of the 79 genes based upon the IQ-TREE model test.

To examine relationships between gene characteristics and levels of concordance/conflict, we calculated the following statistics for each gene: alignment length (a measure of gene information content), tree length (a measure of evolutionary rate), and root-to-tip variance (a measure of variation in evolutionary rate across the tree), all of which can influence properties of phylogenetic inference (Shen, Salichos & Rokas, 2016). This information is presented for each gene across both data types in the supplementary information (Tables S11–S12). We used logistic multiple regression to test relationships between concordance and gene characteristics (Tables S11–S21). We found that alignment length had a significant positive multiplicative relationship with odds of concordance across both datasets (Fig. 2, Tables S14–S17). Tree length and root-to-tip variance had significant positive and negative multiplicative relationships, respectively, in both datasets (Tables S14–S15). Notably, excluding highly influential observations did not affect the direction or significance of most relationships (Tables S16–S17), indicating that the results were not due to the influence of these data points. However, this did increase the estimated impact of alignment length on log odds by a factor of 10, probably due to the high leverage exerted by the very long alignments of *ycf1* and *ycf2*. In models including only alignment length and tree length, both predictors had significant positive multiplicative

relationships with odds of concordance even when the other was included in the model (Fig. 2, Tables S18–S21).

We examined patterns of conflict among individual genes by calculating the pairwise RF distance and producing a network graph (Fig. S4). Most genes had large pairwise RF distances to all other genes. We observed a general cluster of genes that were less distant from one another and from the AT and inferred species trees, including *rpoC2* and *ndhF*, that appeared to correlate with alignment length (Fig. S1). In general, the tree set is characterized by high levels of discordance among gene trees as well as between gene trees and the AT and inferred species trees.

### Genomic patterns of concordance/conflict

In terms of number of nodes concordant with the AT, the LSC and SSC regions were roughly comparable, with both outperforming the IR regardless of whether BS support was considered (Table S13). This pattern held across time periods, with the LSC and SSC regions having more concordant nodes than the IR at every time slice. The tree lengths of the LSC and SSC regions (1.82 and 2.05, respectively) were also considerably larger than that of the IR (0.98). The concatenated gene alignment lengths of the LSC, SSC, and IR were 73,422 bp, 10,395 bp, and 19,314 bp, respectively. The genome diagram of concordance (Fig. S1) does not show any striking patterns among the different genomic regions; however, it is notable that the majority of the LSC region is discordant (or uninformative) except for a few highly informative genes (namely, *rpoC2* and *matK*).

### Performance of individual plastid genes

Across all analyses, *rpoC2* showed the highest levels of concordance with the AT (Fig. 2, Tables S11–S12); in general, when support was not considered, it performed at least as well as the 79 concatenated genes in reconstructing the AT (Fig. 3). The commonly used genes *ndhF* and *matK* generally scored among the best-performing plastid genes (in terms of number of concordant nodes), while *rbcl*, the other most commonly used gene, performed relatively poorly (Fig. 2). The *matK* alignment is ~250 bp longer than the *rbcl* alignment; the best performing gene, *rpoC2* (alignment length 4,660 bp), is one of the longest plastid genes. However, the notably long region *ycf1*—the unaligned sequence of which is ~5,400 bp (Dong et al., 2015)—did not perform as well as *rpoC2*. In this study, the alignment length of *ycf1* was 21,696 bp (vs. 4,660 bp for *rpoC2*). In several cases *ycf1* performed toward the top; however, it was never the top-performing gene in terms of number of concordant nodes (Tables S11–S12). Despite the high levels of observed conflict overall, we found that every node of the AT was supported by at least one gene. Thus, to varying degrees, all relationships of the AT are found within the plastome gene tree set.

### Saturation analyses

None of the genes analyzed showed significant signatures of saturation (Fig. S5). The uncorrected genetic distances and the corrected genetic distances appeared to be roughly the same, and this trend was true for the first, second, and third codon position. Four genes (*ndhD*, *psbL*, *rpl16*, and *rpl2*) were unable to be properly analyzed for saturation as

**Table 1** Log-Likelihood scores rounded to the nearest whole number for alternative resolutions of contentious nodes based on edge-based analyses using a modification of the Maximum Gene Wise Edge (MGWE) method (Walker, Brown & Smith, 2018). The best-supported relationship (highest log-likelihood score) for each case is presented in bold.

| Contentious relationships                                 | GT             | AT             | CC             | Alternative topology (>3 genes supporting) |
|---|----------------|----------------|----------------|--|
| <i>Buxus</i> placement (node 10)                          | <b>-618242</b> | -618371        | -618371        | <b>-618242</b>                             |
| Lamiid relationships (node 19)                            | <b>-618203</b> | -618235        | <b>-618203</b> | <b>-618203</b>                             |
| <i>Jacobaea</i> and <i>Artemisia</i> placements (node 34) | -618535        | <b>-618506</b> | N/A (BS <70)   | N/A  |
| <i>Galinsoga</i> placement (node 36)                      | NA             | -618308        | <b>-618174</b> | -618319                                    |

**Table 2** Contentious relationships tested using edge-based analyses. Individual genes supporting, with  $\geq 70$ , BS the alternative topologies examined for each contentious relationship.

| Contentious relationships                                 | GT  | AT                                   | CC  | Alternative topology (>3 genes supporting)  |
|---|---|--------------------------------------|---|---|
| <i>Buxus</i> placement (node 10)                          | <i>rbcL, petL, rps2, rps14</i>  | <i>ndhD, rps4, matK, psbB, rpoC2</i> | <i>ndhD, rps4, matK, psbB, rpoC2</i>  | <i>rbcL, petL, rps2, rps14</i>  |
| Lamiid relationships (node 19)                            | <i>ndhK, rbcL, rps11, rps7, ycf4, rpoC2, rps4, ndhA, ndhJ, ccsA, rps2, petB, matK, ndhI, atpB</i> | <i>psbB, petA</i>                    | <i>ndhK, rbcL, rps11, rps7, ycf4, rpoC2, rps4, ndhA, ndhJ, ccsA, rps2, petB, matK, ndhI, atpB</i> | <i>ndhK, rbcL, rps11, rps7, ycf4, rpoC2, rps4, ndhA, ndhJ, ccsA, rps2, petB, matK, ndhI, atpB</i> |
| <i>Jacobaea</i> and <i>Artemisia</i> placements (node 34) | <i>rpoC2</i>  | None                                 | N/A   | N/A   |

the nucleotide data uploaded to GenBank were missing the codons for 2 to 3 amino acids in some taxa resulting in an improper codon alignment.

### Analysis of well-supported conflict

We found that, out of the four major contentious relationships in the AT, the edge-based analysis only supports one (Table 1) compared to the alternative hypotheses examined (i.e., the GT, the plastome trees, and the gene trees). In the case of *Buxus* (node 1 in Fig. 1), this analysis supports *Buxus* as sister to Trochodendraceae (Table 1), which is the topology found in the GT and four gene trees: *rbcL, petL, rps2, rps14* (Table 2). In the case of lamiid relationships (node 2), the data support *Olea* (Lamiales) sister to *Acnistus* (Solanales) + *Coffea* (Gentianales), which is the relationship present in the GT and 15 gene trees (Tables 1 and 2). For node 3 (the placements of *Jacobaea* and *Artemisia*), the data support the AT topology, yet no individual gene trees provided  $\geq 70$  BS support for this relationship; however, *rpoC2* matched the GT topology with strong support (Tables 1 and 2). Lastly, for node 4 (the placement of *Galinsoga*), the data collectively support *Parthenium argentatum* sister to *Galinsoga* (our plastome tree, PT), with two gene trees (*petA, psaJ*) providing strong support for this topology; however, 16 genes (Table 2) supported a dominant alternative (*Ageratina* sister to *Galinsoga quadriradiata*).

## DISCUSSION

Following the typical assumptions of chloroplast inheritance, we would expect all genes in the plastomes to share the same evolutionary history. We would also expect all plastid genes to show similar patterns of conflict when compared to non-plastid inferred phylogenies. Furthermore, the amino acid and nucleotide plastid data we used should show the same conflicting/concordant relationships against given reference trees (e.g., the AT or GT). Our results, however, discussed below, frequently conflict with these common assumptions about chloroplast inheritance and evolutionary history.

### Conflicting topologies inferred from the chloroplast genome

In general, the ‘plastome’ topologies inferred from nucleotide and amino acid alignments showed high levels of concordance with the AT (Fig. 1; Tables S11–S12). While the genes within the chloroplast genome are largely uninformative for most nodes of the phylogeny, a number of genes exhibited well-supported conflict (Fig. 1). In general, there appears to be no relationship between evolutionary scale and amount of gene tree conflict (i.e., conflict generally does not appear to correlate with divergence time in angiosperms). Instead, the extent of conflict/concordance had a stronger relationship with the molecule type analyzed. The amino-acid dataset showed the highest levels of gene tree conflict (Fig. 3, Table S12), and the nucleotide dataset had about half the amount of gene tree conflict found in the amino acid data (Fig. 3, Table S11). Support from edge-based analyses (a topology test that is robust to gene tree conflict at nodes other than the one under consideration) for the GT over the AT suggests that at least some signal for conflicting relationships is derived from underlying gene tree conflict when resolving species relationships using all genes. Positive relationships between alignment length, tree length, and concordance also suggest that stochastic error is at play, because shorter or slower-evolving genes harbor less signal to accurately infer relationships. Pairwise comparisons between chloroplast genes indicated that they did not appear to share the same evolutionary history (Fig. S4). Even longer genes were relatively discordant with one another (for example *rpoC2* and *ycf1* had an RF distance of 27). Against a backdrop of largely uninformative genes, long, influential sequences might drive the results of concatenated analyses (Shen, Salichos & Rokas, 2016). Indeed, *rpoC2* and *ndhF* had small distances to the AT, concatenated topology, and to one another. Because the AT is a synthetic topology based on analyses of all three plant genomes, we do not consider this to be a potential bias (i.e., the AT reflecting the signal of these genes), but instead indicative of our dataset converging on the signal of the true relationships. In our analysis, however, longer genes still had largely discordant trees.

The superior performance of (coding) nucleotide data compared to amino acid data possibly stems from the relatively greater information content of nucleotides (i.e., longer alignments). Assuming there is not a significant amount of missing/indel data (with the exception of *ycf1* and *ycf2*), longer alignments should result in better-informed models of evolution, all else being equal, aided by both parsimony-informative and -uninformative characters (Yang, 1998). However, inherent differences in amino acid and nucleotide models might also explain differences in performance. The nucleotide data were analyzed under the GTR model, where substitution rates among bases are individually estimated; in

contrast, for the empirical amino acid substitution models investigated here, substitution rates are pre-estimated, as the number of estimated parameters for changes among 20 states is extremely large. Additionally, although the plastome has been treated as a single molecule for designing amino acid models of evolution (*Adachi et al., 2000*), a wide variety of amino acid models (some of which were designed for viruses, such as flu or HIV) were inferred to be the best for different plastid genes ([Table S12](#)). This might be the result of the different methods implemented in RAxML vs. IQ-TREE for model testing, the different available models (the best-fitting model for some genes in the RAxML analysis was the unpublished matrix DUMMY2), or the lack of sufficient information (because of gene length) to inform the model. However, the most important point is that, based on the amount of information present in each gene, the chloroplast is inferred to evolve under significantly different models of evolution. Given the highly pectinate structure of the AT, phylogenetic inference in this case should rely heavily on the model for the likelihood calculations. While in some cases (e.g., the shortest plastid genes) limited amounts genetic information might be inherently insufficient, in others, improvements in amino acid modeling might lead to great improvements in phylogenetic inference; such has been suggested for animal mitochondrial data (*Richards et al., 2018*). At least insofar as the best-fitting model adequately models the dynamics of sequence evolution, conflict would be expected to decrease relative to a sub-optimal model if systematic error (through inadequate modeling) was the main driver of our results. However, we found that our results were recapitulated when analyzed using only the WAG model, suggesting that this was not the case. These analyses are not exhaustive, because they can only testify to the impact of systematic error resulting from the variation among the model set considered here and do not consider the adequacy of these models. Even though two genes analyzed under two different models produce the same tree, there could still be common assumptions underlying both models that are violated by the dataset, leading to systematic error. Additionally, two genes may share the best-fitting model and produce the same tree, but that model may be misspecified, leading to error. Although not the focus of our study, future examinations of chloroplast phylogenomics could benefit from tests of model adequacy (*Brown, 2014; Duchêne, Duchêne & Ho, 2018; Chen et al., 2019*).

We expect that at deeper time scales, nucleotides (of coding regions) may begin to experience saturation and thus information loss due to increased noise, at which point amino acids (with 20 states) would begin to outperform nucleotides. However, the time scale of angiosperm evolution does not appear great enough to result in nucleotide saturation (at least for the genes sampled here, given that no genes appeared to exhibit significant levels of saturation; [Fig. S5](#)), indicating that nucleotides outperform amino acids for phylogenetic analysis of plastomes across angiosperms. Future work, with a broader plastome sampling across green plants, will be necessary to determine the evolutionary scale at which amino acids become more informative than nucleotides for phylogenetic inference.

Taxon sampling may also affect the accuracy of phylogenetic inference, and it is possible that some of the incongruence observed here is due to the influence of missing taxa (*Zwickl & Hillis, 2002; Heath, Hedtke & Hillis, 2008*). There is debate about the relative impact of



more sequences versus more taxa ([Rosenberg & Kumar, 2001](#); [Zwickl & Hillis, 2002](#); [Rokas & Carroll, 2005](#)). However, our dataset sampled most major angiosperm lineages, and the total analysis was highly concordant with the AT and the GT. Furthermore, patterns of conflict in this dataset reflect those detected in more targeted sampling efforts ([Foster, Henwood & Ho, 2018](#)). Taken together, this suggests that taxon sampling did not play a large role in our results.

### Analyses of well-supported conflict

The use of a reference topology (AT) afforded us the ability to examine how individual genes within the plastomes agree/conflict with our current, generally accepted hypothesis of angiosperm phylogeny. However, several contentious areas of angiosperm phylogeny remain unresolved, with different topologies recovered across different analyses (e.g., the AT, GT, and PT show small differences in several parts of the tree). Thus, we used edge-based analyses to compare these alternative topologies.

The placement of *Buxus* (Buxales) is a salient example, with different analyses resolving different placements. Our edge-based analyses found *Buxus* sister to *Trochodendron*/Trochdendrales (i.e., the GT topology) to have the highest likelihood score ([Table 1](#)), and this placement was recovered in four plastid genes (*rbcl*, *petL*, *rps2*, *rps14*) with strong support. However, five plastid genes recovered, with strong support, *Buxus* and *Trochodendron* successively sister to the core eudicots (i.e., the AT topology): *ndhD*, *rps4*, *matK*, *psbB*, and *rpoC2*. Relationships among core lamiid orders are another intriguing example of gene tree conflict. Core lamiid relationships have been notoriously problematic ([Refulio-Rodriguez & Olmstead, 2014](#); [Stull et al., 2015](#)), and previous phylogenetic studies have recovered every possible combination of the constituent orders (Boraginales, Gentianales, Lamiales, Solanales). While our sampling does not include Boraginaceae, we tested alternative relationships among the three remaining orders. The relationship (Lamiales(Gentianales,Solanales)), which is present in the GT, received the highest likelihood support and is found in 15 gene trees ([Tables 1 and 2](#)). However, the alternatives were each recovered with strong support by two plastid genes: *psaA* and *atpH* (Gentianales sister to the rest), and *psbB* and *petA* (Solanales sister to the rest). Asteraceae also exhibited several examples of strongly supported conflict ([Tables 1 and 2](#)), but sampling differences across the trees examined make comparisons more difficult.

It is difficult to determine the causes of the observed instances of strongly supported conflict, which can be found at most nodes in the phylogeny ([Fig. 1, Tables S11–S12](#)). Based on the present analyses, we cannot conclude whether this conflict is biological (e.g., genes with different evolutionary histories in a single genome) or a consequence of systematic, stochastic, or other analytical errors (e.g., modeling error, lack of information or misalignment). We have endeavored to control for these factors. For example, it is possible that alignment errors could be responsible for instances of strongly supported conflict ([Richards et al., 2018](#)). However, amongst genes with strongly supported conflict, no obvious misalignment was present, and recapitulation of the results when testing multiple alignment methods and a cleaning procedure suggests misalignment, if present, does not explain the patterns observed here. Although the findings of [Richards et al.](#)

(2018) are supported by a lack of evidence for recombination in animal mitochondria (White *et al.*, 2008), there are reasons to believe that biological gene tree conflict could occur in the chloroplast genome (Sancho *et al.*, 2018), but we stress that we make no attempt to fully characterize this. In parallel to our work, Gonçalves *et al.* (2019) have analyzed the prevalence of conflict across the plastome and have advocated for the use of both concatenation and multispecies coalescent methods (MSC) for inferring full plastome relationships. Our own analysis of the data with the summary coalescent method ASTRAL (Zhang *et al.*, 2018) demonstrates that the different methods result in different relationships (Fig. 1). Specifically, we found much of the conflict co-occurs with conflict from the full plastome supermatrices, with the exception of the generic level resolutions of *Diplostephium*. We note that a core assumption of the MSC is the lack of linkage amongst loci, likely violated by the chloroplast in all but a small number of cases (Edwards *et al.*, 2016; Gonçalves *et al.*, 2019). See Gonçalves *et al.* (2019) for an overview of coalescent methods applied to plastome data. With our data, no analyses we performed fully explained the level of conflict observed and we suggest that the possibility of biological conflict deserves further exploration, especially given the recent documented instances of inter-plastome recombination (Sullivan *et al.*, 2017; Sancho *et al.*, 2018) and exchange between the chloroplast and other genomes (Martin *et al.*, 1998; Martin, 2003; Straub *et al.*, 2013).

### Implication for chloroplast phylogenomics

These examples above are not meant to represent an exhaustive examination of conflicting relationships within angiosperms. Instead, they are intended to highlight several instances of strongly supported conflict within the plastome. These results nevertheless have important implications for plastid phylogenomics, suggesting that concatenated analyses should be performed with caution given the notable presence of gene-tree conflict. The majority of nodes show at least one strongly supported conflicting gene, and many nodes have roughly equivalent numbers of conflicting and concordant genes against the reference tree (Fig. 1). Because of the mode of inheritance of the chloroplast, it has been common for authors to assume that gene tree incongruence occurs due to stochastic error, and thus perform concatenation analyses on the whole chloroplast to increase phylogenetic signal. One of the key assumptions of concatenation methods is that all genes share the same underlying tree. While the signal of stochastic error (e.g., longer genes being more concordant) supports this approach, our detection of supported conflict suggests that this approach may be less valid than previously assumed. Concatenated analyses with extensive underlying conflict can yield problematic results in terms of both topology and branch lengths (Brown & Thomson, 2017; Shen, Hittinger & Rokas, 2017; Walker, Brown & Smith, 2018). Particularly against a backdrop of largely uninformative genes, a few genes exhibiting conflict due to stochastic error, inadequate modeling, or biological sources can drive the analysis toward the wrong topology (Brown & Thomson, 2017; Shen, Hittinger & Rokas, 2017; Walker, Brown & Smith, 2018). We suggest that authors should proceed with caution when assuming that all chloroplast genes share the same tree.

## Utility of individual plastid genes for future studies

Previous studies have laid a strong framework for determining the utility of chloroplast regions at various phylogenetic scales. For example, work by *Shaw et al. (2005)*, *Shaw et al. (2007)* and *Shaw et al. (2014)* highlighted non-coding DNA regions useful for shallow evolutionary studies, while *Graham & Olmstead (2000)* explored protein-coding genes useful for reconstructing deep relationships in angiosperms. Here, we expand upon previous work by using a novel phylogenomic approach, allowing us examining the concordance of individual protein-coding plastid genes with all nodes of the accepted angiosperm phylogeny (AT; as well as slight variations thereof, e.g., the GT). We paid special attention to *matK* and *rbcL*, given their historical significance for plant systematics (e.g., *Donoghue et al., 1992*; *Chase et al., 1993*; *Hilu et al., 2003*). We find that *rbcL* performs relatively well in recapitulating the AT—however, *matK* performs considerably better (i.e., it generally has more nodes concordant with the AT; [Tables S11–S12](#)). This is likely due to a strong positive correlation between alignment length and number of concordant nodes, as noted above ([Fig. 2](#)); *matK* has a longer alignment/gene length than *rbcL*.

The gene *ycf1* has been found to be a useful marker in phylogenetics (e.g., *Neubig et al., 2009*; *Neubig & Abbott, 2010*; *Thomson, Vargas & Dick, 2018*) and barcoding (*Dong et al., 2015*), and here we find that it generally performs above average. The alignment of *ycf1* is abnormally long and contains a large number of gaps, and this is likely due to its position spanning the boundary of the IR and the SSC, an area known to fluctuate greatly in size. This variability likely contributes to the value of *ycf1* as a marker for ‘species-level’ phylogenetics and barcoding. However, the performance of *ycf1* does not scale with its alignment length. In terms of concordance, we find that *matK* performs roughly equally as well if not slightly better than *ycf1* ([Tables S11–S12](#)). This might in part be a consequence of *ycf1* being missing/lacking annotation from some species, preventing us from analyzing its concordance/conflict with certain nodes. Nevertheless, our results add to the body of evidence supporting *ycf1* as a generally useful plastid region. However, we found *rpoC2* to outperform all other plastid regions in every case ([Tables S11–S12](#)), and its alignment length (4,660 bp) is  $\sim 1/5$ th the length of *ycf1*, easing the computational burden of using alignment tools such as FSA and BAli-Phy (which would struggle with a region as long as *ycf1*), as well as divergence dating and tree-building programs such as BEAST (*Suchard & Redelings, 2006*; *Drummond & Rambaut, 2007*).

In our analyses, when BS support is not considered, *rpoC2* performed at least as well if not better than using the concatenation of all chloroplast genes ([Fig. 3](#); [Tables S11–S12](#)). When support is considered ([Fig. 3](#); [Tables S11–S12](#)), *rpoC2* still remains the best-performing gene, but it performs slightly worse than the concatenation of all chloroplast genes (in terms of number of supported nodes concordant with the AT). The utility of *rpoC2* likely stems from its notable length, resulting in a wealth of useful phylogenetic information. Biologically, even long genes may be functionally constrained such that not all sites are variable, such as *rbcL*, which performed relatively poorly here (*Kellogg & Juliano, 1997*). However, were this to be the case for *rpoC2*, its longer length relative to *rbcL* should be expected to counter this effect. In light of our results, *rpoC2* should be a highly attractive coding region for future studies, as it generally recapitulates the plastome phylogeny while allowing more proper

branch length inferences (given that conflicting signal among multiple genes can result in problematic branch length estimates: [Mendes & Hahn, 2016](#)). Recent work ([Salichos & Rokas, 2013](#); [Smith, Brown & Walker, 2018](#); [Steenwyk et al., 2018](#)) indicates that filtering for a smaller set of highly informative genes might yield more accurate results in various phylogenetic applications (e.g., topology estimation, divergence dating). Given this, *rpoC2* might be particularly useful for comparative analyses requiring accurate branch length estimates (i.e., the majority of comparative analyses). Use of *rpoC2* alone (instead of the entire plastome) would also allow for more complex, computationally expensive models to be implemented. Furthermore, focused sequencing of *rpoC2* would increase compatibility of datasets from different studies, facilitating subsequent comprehensive, synthetic analyses. Although the performance *rpoC2* may be dataset dependent, our results support its utility at multiple levels, in terms of both time scales and sampling. It is important to note, however, that we did not include non-coding regions in our study, which would likely outperform coding regions at shallow phylogenetic levels. As previously noted, however, it is generally too difficult to use non-coding regions across broad phylogenetic scales (e.g., angiosperms) because of complications related to homology assessment and alignment.

### Genomic patterns of concordance/conflict

Several previous studies (e.g., [Jian et al., 2008](#); [Moore et al., 2011](#)) have highlighted the Inverted Repeat (IR) as a valuable plastid region for deep-level phylogenetic analyses, attributing its utility to its relatively slow rate of evolution, resulting in less homoplasy and minimal saturation. Our results contradict this notion, suggesting that the coding sequences of the IR alone perform poorly compared to the LSC and SSC coding regions for reconstructing angiosperm phylogeny ([Fig. S1](#) and [Table S13](#)). It is important to note that there are important differences between our study and earlier studies on the IR (e.g., [Jian et al., 2008](#); [Moore et al., 2011](#)). For one, we did not include the ribosomal RNA genes, which are highly conserved; thus if the conserved nature of the IR (or at least portions of it) is the basis of its utility, then this might explain the poor performance in the current study. Our saturation analyses, however, (described above) did not reveal any genes to have significant saturation issues at the scale of angiosperm evolution. This calls into question the idea that the conserved genes of the IR would make it superior for reconstruction of angiosperm phylogeny. Instead, it is possible that the non-coding regions of the IR (which we did not include here) are highly informative for angiosperm phylogeny. While the non-coding regions of the LSC and SSC regions have been extensively examined for use as phylogenetic markers ([Shaw et al., 2005](#); [Shaw et al., 2007](#); [Shaw et al., 2014](#)), the non-coding regions of the IR have been underexplored. Among the IR genes examined here, *ycf2* (which is exceptionally long) showed the greatest levels of concordance ([Fig. S1](#) and [Tables S11–S12](#)), underscoring the idea that longer genes are generally more useful for phylogeny reconstruction.

Another important difference between our study and [Moore et al. \(2011\)](#) is that we only partitioned our data by gene region, while [Moore et al. \(2011\)](#) explored various partitioning strategies (including codon positions and different combinations of genes). It is clear that sequences within the plastomes follow various different models of molecular evolution (as

shown above in the results section *Patterns of conflicting chloroplast signal*). Exploration of more complicated partitioning schemes—which is a time-consuming process (Lanfear *et al.*, 2012; Kainer & Lanfear, 2015), and beyond the scope of this study—might generally improve plastome-inferred phylogenies.

## Conclusions

We find notable levels of gene-tree conflict within the plastome, at all levels of angiosperm phylogeny, highlighting the necessity of future research into the causes of plastome conflict: do some genes share different evolutionary histories, or is systematic error (e.g., poor modeling) the source of the observed conflict? With rare exceptions, most genes are largely uninformative and, when analyzed as a whole, are therefore unlikely to mislead researchers. However, our results call into question the appropriateness of concatenating plastid genes for phylogeny reconstruction without some level of scrutiny. Even if the concatenated topology is largely congruent with the ‘true’ species tree of a given group, underlying conflict can mislead branch length estimation (Mendes & Hahn, 2016). Therefore, we recommend that comparative methods and/or dating analyses that rely on branch lengths use filtering procedures similar to those proposed for nuclear data (Salichos & Rokas, 2013; Smith, Brown & Walker, 2018; Steenwyk *et al.*, 2018).

We find alignment length and molecule type (nucleotide vs amino acids) to be the strongest correlate with plastid gene performance—i.e., longer genes and nucleotides (which have greater information content than amino acids) generally show higher levels of concordance with our accepted hypothesis of angiosperm phylogeny. This also shows that as a single unit, *matK*, *rbcL* and the other commonly used marker *ndhF* are not up to the task alone and if possible should be replaced by longer genes such as *rpoC2*, *ycf1*, and *ycf2*. Our findings support the notion that longer genes (with greater information content) are generally superior for phylogeny reconstruction, in line with theoretical predictions (Yang, 1998) and should be prioritized when developing phylogenetic studies.

## ACKNOWLEDGEMENTS

We thank S.A. Smith and M.J. Moore for reading an early draft of the manuscript and proving helpful suggestions toward its improvement. We also thank A. Alverson, Liming Cai, two anonymous reviewers and S. Edwards for helpful comments that improved the manuscript.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

Joseph F. Walker was supported by a Rackham Predoctoral Fellowship. Gregory W. Stull was supported by a NSF Postdoctoral Fellowship (NSF DBI grant 1612032). Oscar M. Vargas was supported by NSF FESD 1338694 and DEB 1240869. Drew A. Larson was supported by NSF DEB grant 1458466. Nathanael Walker-Hale was supported by a Woolf Fisher Cambridge Scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:

NSF Postdoctoral Fellowship: 1612032.

NSF FESD: 1338694.

DEB: 1240869.

NSF DEB: 1458466.

Woolf Fisher Cambridge Scholarship.

### Competing Interests

The authors declare there are no competing interests.

### Author Contributions

- Joseph F. Walker and Nathanael Walker-Hale conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Oscar M. Vargas and Drew A. Larson conceived and designed the experiments, analyzed the data, approved the final draft.
- Gregory W. Stull conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, approved the final draft.

### Data Availability

The following information was supplied regarding data availability:

Data and code are available at GitHub:

<https://github.com/jfwalker/ChloroplastPhylogenomics>.

### Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.7747#supplemental-information>.

## REFERENCES

- Adachi J, Waddell PJ, Martin W, Hasegawa M. 2000.** Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution* **50**:348–358 DOI [10.1007/s002399910038](https://doi.org/10.1007/s002399910038).
- Bendich AJ. 2004.** Circular chloroplast chromosomes: the grand illusion. *The Plant Cell* **16**:1661–1666 DOI [10.1105/tpc.160771](https://doi.org/10.1105/tpc.160771).
- Birky Jr CW. 1995.** Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences of the United States of America* **92**:11331–11338 DOI [10.1073/pnas.92.25.11331](https://doi.org/10.1073/pnas.92.25.11331).
- Bouillé M, Senneville S, Bousquet J. 2011.** Discordant mtDNA and cpDNA phylogenies indicate geographic speciation and reticulation as driving factors for the diversification of the genus *Picea*. *Tree Genetics & Genomes* **7**(3):469–484.

- Boyce K, Sievers F, Higgins DG. 2015.** Instability in progressive multiple sequence alignment algorithms. *Algorithms for Molecular Biology* **10**(1):Article 26 DOI [10.1186/s13015-015-0057-1](https://doi.org/10.1186/s13015-015-0057-1).
- Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009.** Fast statistical alignment. *PLOS Computational Biology* **5**:e1000392 DOI [10.1371/journal.pcbi.1000392](https://doi.org/10.1371/journal.pcbi.1000392).
- Brown JM. 2014.** Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Systematic Biology* **63**:334–348 DOI [10.1093/sysbio/syu002](https://doi.org/10.1093/sysbio/syu002).
- Brown JM, Thomson RC. 2017.** Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology* **66**:517–530 DOI [10.1093/sysbio/syw101](https://doi.org/10.1093/sysbio/syw101).
- Brown JW, Walker JF, Smith SA. 2017.** Phyx: phylogenetic tools for unix. *Bioinformatics* **33**:1886–1888 DOI [10.1093/bioinformatics/btx063](https://doi.org/10.1093/bioinformatics/btx063).
- Burleigh JG, Mathews S. 2007a.** Assessing among-locus variation in the inference of seed plant phylogeny. *International Journal of Plant Sciences* **168**:111–124 DOI [10.1086/509586](https://doi.org/10.1086/509586).
- Burleigh JG, Mathews S. 2007b.** Assessing systematic error in the inference of seed plant phylogeny. *International Journal of Plant Sciences* **168**:125–135 DOI [10.1086/509588](https://doi.org/10.1086/509588).
- Carbonell-Caballero J, Alonso R, Ibañez V, Terol J, Talon M, Dopazo J. 2015.** A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus citrus. *Molecular Biology and Evolution* **32**:2015–2035 DOI [10.1093/molbev/msv082](https://doi.org/10.1093/molbev/msv082).
- Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu YL, Kron KA. 1993.** Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. *Annals of the Missouri Botanical Garden* **1**:528–580 DOI [10.2307/2399846](https://doi.org/10.2307/2399846).
- Chatzou M, Floden EW, Di Tommaso P, Gascuel O, Notredame C. 2018.** Generalized bootstrap supports for phylogenetic analyses of protein sequences incorporating alignment uncertainty. *Systematic Biology* **67**(6):997–1009.
- Chen W, Kenney T, Bielawski J, Gu H. 2019.** Testing adequacy for DNA substitution models. *BMC Bioinformatics* **20**:349 DOI [10.1186/s12859-019-2905-3](https://doi.org/10.1186/s12859-019-2905-3).
- Corriveau JL, Coleman AW. 1988.** Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperms. *American Journal of Botany* **75**:1443–1458.
- Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J. 2012.** Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany* **99**:291–311 DOI [10.3732/ajb.1100356](https://doi.org/10.3732/ajb.1100356).
- Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008.** Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* **36**:e122–e122 DOI [10.1093/nar/gkn502](https://doi.org/10.1093/nar/gkn502).
- Csardi G, Nepusz T. 2006.** The igraph software package for complex network research. *Interjournal Complex Systems* **1695**(5):1–9.

- Davis JJ, Mcneal JR, Barrett CF, Chase MW, Cohen JJ, Duvall MR, Givnish TJ, Graham SW, Petersen G, Chris Pires J, Seberg O, Stevenson DW, Leebens-Mack J. 2013.** Contrasting patterns of support among plastid genes and genomes for major clades of the monocotyledons. Chapter. In: Wilkin P, Mayo SJ, eds. *Early events in monocot evolution. Systematics association special volume series*, Cambridge: Cambridge University Press.
- Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S. 2015.** *ycf1*, the most promising plastid DNA barcode of land plants. *Scientific Reports* 5:8348 DOI 10.1038/srep08348.
- Donoghue MJ, Olmstead RG, Smith JF, Palmer JD. 1992.** Phylogenetic relationships of Dipsacales based on *rbcL* sequences. *Annals of the Missouri Botanical Garden* 79:333–345 DOI 10.2307/2399772.
- Doyle JJ. 1992.** Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* 17:144–163 DOI 10.2307/2419070.
- Drummond AJ, Rambaut A. 2007.** BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7:214 DOI 10.1186/1471-2148-7-214.
- Duchêne DA, Duchêne S, Ho SYW. 2018.** Differences in performance among test statistics for assessing phylogenomic model adequacy. *Genome Biology and Evolution* 10:1375–1388 DOI 10.1093/gbe/evy094.
- Erixon P, Oxelman B. 2008.** Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast *clpP1* gene. *PLOS ONE* 3(1):e1386.
- Felsenstein J, Churchill GA. 1996.** A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* 13(1):93–104 DOI 10.1093/oxfordjournals.molbev.a025575.
- Foster CS, Henwood MJ, Ho SY. 2018.** Plastome sequences and exploration of tree-space help to resolve the phylogeny of riceflowers (Thymelaeaceae: Pimelea). *Molecular Phylogenetics and Evolution* 127:156–167.
- Fruchterman TMJ, Reingold EM. 1991.** Graph drawing by force-directed placement. *Software: Practice and Experience* 21:1129–1164.
- Galtier N, Daubin V. 2008.** Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363:4023–4029 DOI 10.1098/rstb.2008.0144.
- Gitzendanner MA, Soltis PS, Wong GK, Ruhfel BR, Soltis DE. 2018.** Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *American Journal of Botany* 105:291–301 DOI 10.1002/ajb2.1048.
- Gonçalves DJP, Simpson BB, Ortiz EM, Shimizu GH, Jansen RK. 2019.** Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. *Molecular Phylogenetics and Evolution* 138:219–232 DOI 10.1016/j.ympev.2019.05.022.
- Graham SW, Olmstead RG. 2000.** Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *American Journal of Botany* 87:1712–1730 DOI 10.2307/2656749.



- Heath TA, Hedtke SM, Hillis DM. 2008.** Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematic and Evolution* **46**(3):239–257.
- Hansen AK, Escobar LK, Gilbert LE, Jansen RK. 2007.** Paternal, maternal, and bipaternal inheritance of the chloroplast genome in *Passiflora* (Passifloraceae): implications for phylogenetic studies. *American Journal of Botany* **94**(1):42–46 DOI [10.3732/ajb.94.1.42](https://doi.org/10.3732/ajb.94.1.42).
- Hillis DM, Bull JJ. 1993.** An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**:182–192 DOI [10.1093/sysbio/42.2.182](https://doi.org/10.1093/sysbio/42.2.182).
- Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R. 2003.** Angiosperm phylogeny based on *matK* sequence information. *American Journal of Botany* **90**:1758–1776 DOI [10.3732/ajb.90.12.1758](https://doi.org/10.3732/ajb.90.12.1758).
- Hlavac M. 2018.** Stargazer: well-formatted regression and summary statistics tables. R package version 5.2.1. Available at <https://CRAN.R-project.org/package=stargazer>.
- Huang S, Chiang YC, Schaal BA, Chou CH, Chiang TY. 2001.** Organelle DNA phylogeography of *Cycas taitungensis*, a relict species in Taiwan. *Molecular Ecology* **10**(11):2669–2681.
- Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK. 2007.** Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences* **104**:19369–19374 DOI [10.1073/pnas.0709121104](https://doi.org/10.1073/pnas.0709121104).
- Jian S, Soltis PS, Gitzendanner MA, Moore MJ, Li R, Hendry TA, Qiu YL, Dhingra A, Bell CD, Soltis DE. 2008.** Resolving an ancient, rapid radiation in Saxifragales. *Systematic Biology* **57**:38–57 DOI [10.1080/10635150801888871](https://doi.org/10.1080/10635150801888871).
- Johnson LB, Palmer JD. 1988.** Heteroplasmy of chloroplast DNA in *Medicago*. *Plant Molecular Biology* **12**:3–11.
- Kainer D, Lanfear R. 2015.** The effects of partitioning on phylogenetic inference. *Molecular Biology and Evolution* **32**:1611–1627 DOI [10.1093/molbev/msv026](https://doi.org/10.1093/molbev/msv026).
- Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017.** ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**:587–589 DOI [10.1038/nmeth.4285](https://doi.org/10.1038/nmeth.4285).
- Katoh K, Standley DM. 2013.** MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**(4):772–780 DOI [10.1093/molbev/mst010](https://doi.org/10.1093/molbev/mst010).
- Kellogg EA, Juliano ND. 1997.** The structure and function of RuBisCO and their implications for systematic studies. *American Journal of Botany* **84**(3):413–428 DOI [10.2307/2446015](https://doi.org/10.2307/2446015).
- Lanfear R, Calcott B, Ho SY, Guindon S. 2012.** PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* **29**:1695–1701 DOI [10.1093/molbev/mss020](https://doi.org/10.1093/molbev/mss020).

- Lee DJ, Blake TK, Smith SE. 1988. Biparental inheritance of chloroplast DNA and the existence of heteroplasmic cells in alfalfa. *Theoretical and Applied Genetics* 76:545–549 DOI 10.1007/BF00260905.
- Li H-T, Yi T-S, Gao L-M, Ma P-F, Zhang T, Yang J-B, Gitzendanner MA, Fritsch PW, Cai J, Luo Y, Wang H, Van der Bank M, Zhang S-D, Wang Q-F, Wang J, Zhang Z-R, Fu C-N, Yang J, Hollingsworth PM, Chase MW, Soltis DE, Soltis PS, Li D-Z. 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants* 5(5):461–470 DOI 10.1038/s41477-019-0421-0.
- Liu J, Qi ZC, Zhao YP, Fu CX, Jenny Xiang QY. 2012. Complete cpDNA genome sequence of Smilax china and phylogenetic placement of Liliales—influences of gene partitions and taxon sampling. *Molecular Phylogenetics and Evolution* 64:545–562 DOI 10.1016/j.ympev.2012.05.010.
- Logacheva MD, Penin AA, Samigullin TH, Vallejo-Roman CM, Antonov AS. 2007. Phylogeny of flowering plants by the chloroplast genome sequences: in search of a lucky gene. *Biochemistry* 72(12):1324–1330.
- Maddison W. 1997. Gene trees in species trees. *Systematic Biology* 46:523–536 DOI 10.1093/sysbio/46.3.523.
- Magallón S, Gómez-Acevedo S, Sánchez-Reyes L, Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist* 207:437–453 DOI 10.1111/nph.13264.
- Marshall HD, Newton C, Ritland K. 2001. Sequence-repeat polymorphisms exhibit the signature of recombination in lodgepole pine chloroplast DNA. *Molecular Biology and Evolution* 18(11):2136–2138.
- Martin W. 2003. Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proceedings of the National Academy of Sciences of the United States of America* 100:8612–8614 DOI 10.1073/pnas.1633606100.
- Martin W, Stoebe B, Goremykin V, Hapsmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393:162–165 DOI 10.1038/30234.
- McCauley DE, Sundby AK, Bailey MF, Welch ME. 2007. Inheritance of chloroplast DNA is not strictly maternal in *Silene vulgaris* (Caryophyllaceae): evidence from experimental crosses and natural populations. *American Journal of Botany* 94:1333–1337 DOI 10.3732/ajb.94.8.1333.
- Mendes FK, Hahn MH. 2016. Gene tree discordance causes apparent substitution rate variation. *Systematic Biology* 65:711–721 DOI 10.1093/sysbio/syw018.
- Mogensen HL. 1996. The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany* 83:383–404 DOI 10.1002/j.1537-2197.1996.tb12718.x.
- Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences of the United States of America* 104:19363–19368 DOI 10.1073/pnas.0708072104.

- Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Foltá KM, Soltis DE. 2006.** Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology* **6**:17 DOI [10.1186/1471-2229-6-17](https://doi.org/10.1186/1471-2229-6-17).
- Moore MJ, Hassan N, Gitzendanner MA, Bruenn RA, Croley M, Vandeventer A, Horn JW, Dhingra A, Brockington SF, Latvis M. 2011.** Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. *International Journal of Plant Sciences* **172**:541–558 DOI [10.1086/658923](https://doi.org/10.1086/658923).
- Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010.** Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences of the United States of America* **107**:4623–4628 DOI [10.1073/pnas.0907801107](https://doi.org/10.1073/pnas.0907801107).
- Morrison DA. 2009.** Why would phylogeneticists ignore computerized sequence alignment? *Systematic Biology* **58**(1):150–158 DOI [10.1093/sysbio/syp009](https://doi.org/10.1093/sysbio/syp009).
- Morrison DA, Morgan MJ, Kelchner SA. 2015.** Molecular homology and multiple-sequence alignment: an analysis of concepts and practice. *Australian Systematic Botany* **28**(1):46–62 DOI [10.1071/SB15001](https://doi.org/10.1071/SB15001).
- Neubig KM, Abbott JR. 2010.** Primer development for the plastid region *ycf1* in Annonaceae and other magnoliids. *American Journal of Botany* **97**:e52–e55 DOI [10.3732/ajb.1000128](https://doi.org/10.3732/ajb.1000128).
- Neubig KM, Whitten WM, Carlswald BS, Blanco MA, Endara L, Williams NH, Moore MJ. 2009.** Phylogenetic utility of *ycf1* in orchids: a plastid gene more variable than *matK*. *Plant Systematics and Evolution* **277**:75–84 DOI [10.1007/s00606-008-0105-0](https://doi.org/10.1007/s00606-008-0105-0).
- Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2014.** IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**:268–274.
- Palmer JD. 1983.** Chloroplast DNA exists in two orientations. *Nature* **301**:92–93 DOI [10.1038/301092a0](https://doi.org/10.1038/301092a0).
- Palmer JD. 1985.** Comparative organization of chloroplast genomes. *Annual Review of Genetics* **19**:325–354 DOI [10.1146/annurev.ge.19.120185.001545](https://doi.org/10.1146/annurev.ge.19.120185.001545).
- Paradis E, Schliep K. 2018.** ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**(3):526–528.
- Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011.** Resolving difficult phylogenetic questions: why more sequences are not enough. *PLOS Biology* **9**:e1000602 DOI [10.1371/journal.pbio.1000602](https://doi.org/10.1371/journal.pbio.1000602).
- Philippe H, Sörhannus U, Baroin A, Perasso R, Gasse F, Adoutte A. 1994.** Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. *Journal of Evolutionary Biology* **7**(2):247–265 DOI [10.1046/j.1420-9101.1994.7020247.x](https://doi.org/10.1046/j.1420-9101.1994.7020247.x).
- R Core Team. 2018.** R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Redelings B. 2014.** Erasing errors due to alignment ambiguity when estimating positive selection. *Molecular Biology and Evolution* **31**:1979–1993 DOI [10.1093/molbev/msu174](https://doi.org/10.1093/molbev/msu174).

- Refulio-Rodriguez NF, Olmstead RG. 2014. Phylogeny of Lamiidae. *American Journal of Botany* 101:287–299 DOI 10.3732/ajb.1300394.
- Rice DW, Alverson AJ, Richardson AO, Young GJ, Virginia Sanchez-Puerta M, Munzinger J, Barry K, Boore JL, Zhang Y, DePamphilis CW, Knox EB, Palmer JD. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science* 342(6165):1468–1473 DOI 10.1126/science.1246275.
- Richards EJ, Brown JM, Barley AJ, Chong RA, Thomson RC. 2018. Variation across mitochondrial gene trees provides evidence for systematic error: how much gene tree variation is biological? *Systematic Biology* 67(5):847–860 DOI 10.1093/sysbio/syy013.
- Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53:131–147 DOI 10.1016/0025-5564(81)90043-2.
- Rodríguez-Ezpeleta N, Brinkmann H, Burger G, Roder AJ, Gray MW, Philippe H, Lang BF. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. *Current Biology* 17:1420–1425 DOI 10.1016/j.cub.2007.07.036.
- Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution* 22(5):1337–1344 DOI 10.1093/molbev/msi121.
- Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804 DOI 10.1038/nature02053.
- Roquet C, Sanmartín I, Garcia-Jacas N, Sáez L, Susanna A, Wikström N, Aldasoro JJ. 2009. Reconstructing the history of Campanulaceae with a Bayesian approach to molecular dating and dispersal–vicariance analyses. *Molecular Phylogenetics and Evolution* 52:575–587 DOI 10.1016/j.ympev.2009.05.014.
- Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proceedings of the National Academy of Sciences of the United States of America* 98(19):10751–10756 DOI 10.1073/pnas.191248498.
- Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms—inferring the phylogeny of green plants (*Viridiplantae*) from 360 plastid genomes. *BMC Evolutionary Biology* 14:23 DOI 10.1186/1471-2148-14-23.
- Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449):327–331 DOI 10.1038/nature12130.
- Sancho R, Cantalapiedra CP, López-Alvarez D, Gordon SP, Vogel JP, Catalán P, Contreras-Moreira B. 2018. Comparative plastome genomics and phylogenomics of *Brachypodium*: flowering time signatures, introgression and recombination in recently diverged ecotypes. *New Phytologist* 218:1631–1644 DOI 10.1111/nph.14926.
- Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92:142–166 DOI 10.3732/ajb.92.1.142.

- Shaw J, Lickey EB, Schilling EE, Small RL. 2007.** Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94**:275–288 DOI [10.3732/ajb.94.3.275](https://doi.org/10.3732/ajb.94.3.275).
- Shaw J, Shafer HL, Leonard OR, Kovach MJ, Schorr M, Morris AB. 2014.** Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. *American Journal of Botany* **101**:1987–2004 DOI [10.3732/ajb.1400398](https://doi.org/10.3732/ajb.1400398).
- Shen X-X, Hittinger CT, Rokas A. 2017.** Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* **1**:Article 0126.
- Shen X-X, Salichos L, Rokas A. 2016.** A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. *Genome Biology and Evolution* **8**(8):2565–2580 DOI [10.1093/gbe/evw179](https://doi.org/10.1093/gbe/evw179).
- Shepherd LD, Holland BR, Perrie LP. 2008.** Conflict amongst chloroplast DNA sequences obscures the phylogeny of a group of *Asplenium* ferns. *Molecular Phylogenetics and Evolution* **48**:176–187 DOI [10.1016/j.ympev.2008.02.023](https://doi.org/10.1016/j.ympev.2008.02.023).
- Smith SE. 1989.** Biparental inheritance of organelles and its implications in crop improvement. *Plant Breeding Reviews* **6**:361–393.
- Smith DR. 2014.** Mitochondrion-to-plastid DNA transfer: it happens. *New Phytologist* **202**(3):736–738 DOI [10.1111/nph.12704](https://doi.org/10.1111/nph.12704).
- Smith SA, Brown JW, Walker JF. 2018.** So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLOS ONE* **13**(5):e0197433.
- Smith SA, Moore MJ, Brown JW, Yang Y. 2015.** Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* **15**:150 DOI [10.1186/s12862-015-0423-0](https://doi.org/10.1186/s12862-015-0423-0).
- Smith SA, Walker-Hale N, Walker JF, Brown JW. 2018.** Nested phylogenetic conflicts, combinability, and deep phylogenomics in plants. ArXiv preprint. [arXiv:371930](https://arxiv.org/abs/371930) DOI [10.1101/371930](https://doi.org/10.1101/371930).
- Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker Michael JB, Moore Barbara J, Carlswald S, Bell CD, Latvis M, Crawley S, Black C, Diouf D, Xi Z, Rushworth CA, Gitzendanner MA, Sytsma KJ, Qiu Y-L, Hilu KW, Davis CC, Sanderson MJ, Beaman RS, Olmstead RG, Judd WS, Donoghue MJ, Soltis PS. 2011.** Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* **98**(4):704–730 DOI [10.3732/ajb.1000404](https://doi.org/10.3732/ajb.1000404).
- Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoot SB, Fay MF, Axtell M, Swensen SM, Prince LM, John Kress W, Nixon KC, Farris JS. 2000.** Angiosperm phylogeny inferred from 18S rDNA, *rbcl*, and *atpB* sequences. *Botanical Journal of the Linnean Society* **133**(4):381–346 DOI [10.1111/j.1095-8339.2000.tb01588.x](https://doi.org/10.1111/j.1095-8339.2000.tb01588.x).
- Stamatakis A. 2014.** RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313 DOI [10.1093/bioinformatics/btu033](https://doi.org/10.1093/bioinformatics/btu033).

- Steenwyk JL, Shen X-X, Lind AL, Goldman GG, Rokas A. 2018.** A robust phylogenomic timetree for biotechnologically and medically important fungi from Aspergillaceae (Eurotiomycetes, Ascomycota). ArXiv preprint. [arXiv:370429](https://arxiv.org/abs/370429).
- Stegemann S, Hartmann S, Ruf S, Bock R. 2003.** High-frequenting gene transfer from the chloroplast genome to the nucleus. *Proceedings of the National Academy of Sciences* **100**:8828–8833 DOI [10.1073/pnas.1430924100](https://doi.org/10.1073/pnas.1430924100).
- Straub SCK, Cronn RC, Edwards C, Fishbein M, Liston A. 2013.** Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biology and Evolution* **5**:1872–1885 DOI [10.1093/gbe/evt140](https://doi.org/10.1093/gbe/evt140).
- Stull GW, Moore MJ, Mandala VS, Douglas NA, Kates HR, Qi X, Brockington SF, Soltis PS, Soltis DE, Gitzendanner MA. 2013.** A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* **1**:1794–1813 DOI [10.3732/apps.1200497](https://doi.org/10.3732/apps.1200497).
- Stull GW, Duno de Stefano R, Soltis DE, Soltis PS. 2015.** Resolving basal lamiid phylogeny and the circumscription of Icacinaceae with a plastome-scale data set. *American Journal of Botany* **102**:1794–1813 DOI [10.3732/ajb.1500298](https://doi.org/10.3732/ajb.1500298).
- Suchard MA, Redelings BD. 2006.** BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* **22**:2047–2048 DOI [10.1093/bioinformatics/btl175](https://doi.org/10.1093/bioinformatics/btl175).
- Sullivan AR, Schiffthaler B, Thompson SL, Street NR, Wang XR. 2017.** Interspecific plastome recombination reflects ancient reticulate evolution in *Picea* (Pinaceae). *Molecular Biology and Evolution* **34**:1689–1701 DOI [10.1093/molbev/msx111](https://doi.org/10.1093/molbev/msx111).
- Taberlet P, Gielly L, Pautou G, Bouvet J. 1991.** Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* **17**:1105–1109 DOI [10.1007/BF00037152](https://doi.org/10.1007/BF00037152).
- Thomson AM, Vargas OM, Dick CW. 2018.** Complete plastome sequences from *Bertholletia excelsa* and 23 related species yield informative markers for Lecythidaceae. *Applications in Plant Sciences* **6**:e1151.
- Uribe-Convers S, Duke JR, Moore MJ, Tank DC. 2014.** A long PCR-based approach for DNA enrichment prior to next-generation sequencing for systematic studies. *Applications in Plant Sciences* **2**:Article 1300063 DOI [10.3732/apps.1300063](https://doi.org/10.3732/apps.1300063).
- Vargas OM, Ortiz EM, Simpson BB. 2017.** Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: Diplostephium). *New Phytologist* **214**:1736–1750 DOI [10.1111/nph.14530](https://doi.org/10.1111/nph.14530).
- Walker JF, Jansen RK, Zanis MJ, Emery NC. 2015.** Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. *American Journal of Botany* **102**:1751–1752 DOI [10.3732/ajb.1500299](https://doi.org/10.3732/ajb.1500299).
- Walker JF, Yang Y, Moore MJ, Mikenas J, Timoneda A, Brockington SF, Smith SA. 2017.** Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany* **104**:858–867 DOI [10.3732/ajb.1700083](https://doi.org/10.3732/ajb.1700083).

- Walker JF, Zanis MJ, Emery NC. 2014. Comparative analysis of complete chloroplast genome sequence and inversion variation in *Lasthenia burkei* (Madieae, Asteraceae). *American Journal of Botany* **101**:722–729 DOI [10.3732/ajb.1400049](https://doi.org/10.3732/ajb.1400049).
- Walker JF, Brown JW, Smith SA. 2018. Analyzing contentious relationships and outlier genes in phylogenomics. *Systematic Biology* **67**:916–924 DOI [10.1093/sysbio/syy043](https://doi.org/10.1093/sysbio/syy043).
- Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18**(5):691–699 DOI [10.1093/oxfordjournals.molbev.a003851](https://doi.org/10.1093/oxfordjournals.molbev.a003851).
- White DJ, Wolff JN, Pierson M, Gemmell NJ. 2008. Revealing the hidden complexities of mtDNA inheritance. *Molecular Ecology* **17**(23):4925–4942 DOI [10.1111/j.1365-294X.2008.03982.x](https://doi.org/10.1111/j.1365-294X.2008.03982.x).
- Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR, Wafula E, Der JP, Graham SW, Mathews S, Melkonian M, Soltis DE, Soltis PS, Miles NW, Rothfels CJ, Pokorny L, Shaw AJ, DeGironimo L, Stevenson DW, Surek B, Villarreal JC, Roure B, Philippe H, dePamphilis CW, Chen T, Deyholos MK, Baucom RS, Kutchan TM, Augustin MM, Wang J, Zhang Y, Tian Z, Yan Z, Wu X, Sun X, Wong GK-S, Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* **111**(45):e4859–68.
- Wolfe AD, Randle CP. 2004. Recombination, heteroplasmy, haplotype polymorphism, and paralogy in plastid genes: implications for plant molecular systematics. *Systematic Botany* **29**:1011–1020 DOI [10.1600/0363644042451008](https://doi.org/10.1600/0363644042451008).
- Wolfe KH, Li W-H, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America* **84**:9054–9058 DOI [10.1073/pnas.84.24.9054](https://doi.org/10.1073/pnas.84.24.9054).
- Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. *Systematic Biology* **47**:125–133 DOI [10.1080/106351598261067](https://doi.org/10.1080/106351598261067).
- Zeng L, Zhang Q, Sun R, Kong H, Zhang N, Ma H. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature Communications* **5**:Article 4956 DOI [10.1038/ncomms5956](https://doi.org/10.1038/ncomms5956).
- Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**(6):153 DOI [10.1186/s12859-018-2129-y](https://doi.org/10.1186/s12859-018-2129-y).
- Zhang Q, Liu Y, Sodmergen . 2003. Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant and Cell Physiology* **44**:941–951 DOI [10.1093/pcp/pcg121](https://doi.org/10.1093/pcp/pcg121).
- Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* **51**(4):588–598 DOI [10.1080/10635150290102339](https://doi.org/10.1080/10635150290102339).