Novel Phylogenomic Methods for Uncovering the Evolutionary History of the Hyperdiverse Clade Caryophyllales

by

Joseph F. Walker

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in the University of Michigan
2018

Doctoral Committee:

Associate Professor Stephen A. Smith, Co-Chair
Professor Patricia J. Wittkopp, Co-Chair
Associate Professor Timothy Y. James
Associate Professor Daniel L. Rabosky
Professor John Schiefelbein

Joseph F. Walker

jfwalker@umich.edu

ORCID iD: 0000-0003-2928-8899

To my Mom, Dad, and sister, for all your love and support

Acknowledgements

I would like to start by thanking my **sister**, Sheryl Walker, who gave me the wise advice to work in a lab as an undergraduate so that I would have a job during the summers. Her advice and encouragement have led to me being here today. I would also like to thank my **parents, Robin and Brian Walker,** for the integral role they have played in making sure I am able to pursue my dreams. Growing up, my parents always made sure we lived in an area where my sister and I could receive a proper education, and I can never thank them enough for that. I also want to thank my **Grandfather**, **Grandmother**, and **extended family** all of whom who have encouraged me along the way.

I have been lucky to have two high school friends just a train ride away, and for that I'd like to thank **Matt Schertz** and **Eric Steinberg**. I hope my use of their apartment to escape grad school life was repaid by answering their extremely scientific questions such as "how do lemons have water?" at 11pm on a Sunday. I would also like to thank the friends I have made at Michigan, including the **EEB cohort of 2014**, which will always remain the best cohort. **Peter Cerda**, **Camden Gowler, Jon Massey**, and I somehow managed to never get tired of Taste of Suvai. I'd like to thank **Caroline Parins-Fukuchi** for all the help and advice you've given throughout my thesis. I'd also like to thank **Thomas Jenkinson**, **Alex Taylor,** and numerous others in the EEB department for all the advice and great times we have had together.

Continuing with friends, I would like to thank **Nathanael Walker-Hale**, **Matthieu Grange-Guermente,** and of course **Robyn Phillips**. Nat and Matt both made graduate school

less of a stressful event, and more whether it is formal halls or going to hop and grain I can't thank you enough for your friendship. Robyn has been there to see the low points and the high points of my graduate career and supported me through all of them. I cannot thank her enough for always being there. She listened to the moaning and groaning, and has probably heard more about my thesis than anyone. Thank you so much.

I'd like to thank the members of the **Smith lab**, especially the proofreader of much of my thesis and other scientific documents, **Greg Stull**. I'd like to thank **Joseph Brown,** who has been like another advisor to me; I have learned an enormous amount from Joseph and to this day and I am still learning a lot from him. I'd also like to thank **Ya Yang**, **Ning Wang**, **Oscar Vargas**, **James Pease**, **Diego Alvarado Serrano,** and **Cody Hinchliff**. I've learned so much from all the postdocs who I've been lucky enough to spend time with. I'd also like to thank **Drew Larson** and **Lijun Zhao** for being great labmates. Finally, I'd like to thank **Sonia Ahluwalia** who spent four years as an undergraduate working in lab; during this time, she made invaluable contributions to the research presented in this thesis.

I'd like to thank the amazing EEB staff who helped me complete the formal requirements of the graduate program. Cindy Carl and Carol Solomon have helped me fill out countless forms and made sure I did not forget to meet the requirements of the program. I'd also like to thank members of the **Brockington lab**, and especially **Sam** for hosting me, and **Alfonso** for inviting me to everything the day I arrived and making me feel at home. I also want to thank the **Glover lab** for all the fun conversations at the tearoom and **Roisin** and **Chris** for the somehow infinite stories. I'd like to especially thank **Edwige**, who was also nice enough to host me, along with giving helpful advice throughout graduate school.

On the note of advice, I'd like to thank my committee members, Dr. **Jon Scheifelbein**, Dr. **Tim James**, and Dr. **Dan Rabosky**, all of whom have at some point met with me for both career advice and advice on these projects. I'd like to thank my co-chair **Dr. Patricia Wittkopp,** who has always welcomed me into her lab and given advice on how to deal with academia. Tricia has provided encouragement and support to make this all possible, even though my project has found me spending most of my time in the lab of my other advisor, **Dr. Stephen A Smith.**

Which brings me **Stephen**, who I cannot thank enough for his help, patience, and guidance over the past four years. Over my time in graduate school, I have noticed that scientists often reflect their advisor, especially in how they conduct themselves in academia I am therefore grateful to have been trained in Stephen's lab, as I cannot imagine a better scientific role model to follow. Despite being a great scientist, Stephen is also humble, an undervalued trait. Stephen let me sit in on important meetings and treated me like a colleague. He does not ever look down upon someone for not knowing something (I know this from a lot of first-hand experience), and he takes time to help and explain difficult concepts. Without Stephen, this thesis would not have been possible—I offer him my utmost thanks.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

Gene tree conflict has historically been attributed to methodological error, but can also arise from an array of biological processes (e.g. hybridization, incomplete lineage sorting, etc…). Affordable genome sequencing has made it possible to study patterns of gene tree conflict, which in turn provides evidence for the evolutionary events that led to this conflict. This thesis applies and formulates novel methods for analyzing gene tree conflict to better understand the evolutionary history of the hyperdiverse plant clade Caryophyllales. The uniparentally inherited chloroplast genome is used as a basis for understanding the degree of gene tree conflict that arises from systematic error. Despite typically acting as a single locus, whereby all genes share one evolutionary history, I examined a large chloroplast genome dataset that showed substantial gene tree conflict. Conflict was significantly correlated with the length of the genes' alignment and the amount of evolutionary information within the gene. I also examined sources of gene tree conflict with a transcriptome dataset of the carnivorous clade within the Caryophyllales. The carnivorous clade contained evidence of ancient hybridization and high levels of gene/genome duplications. Every family in the clade, except one, had a unique paleopolyploidy event. The most highly debated node of the carnivorous Caryophyllales had a large amount of gene tree conflict, and phylogenetic results were highly sensitive to taxon sampling. Using the largest phylogenomic dataset to date, I also reconstructed the evolutionary relationships among lineages across the entirety of the clade Caryophyllales. This was done using a new computationally tractable and phylogenetically informed method of hierarchical clustering. The results provide an outline of highly conflicting regions in the clade Caryophyllales and provide evidence for a new family, Agdestidaceae, which is proposed based on morphology and phylogenetic position. Furthermore, despite the large amount of data, the inferred phylogeny was highly congruent with the three gene phylogenies found in the literature. This demonstrates that our current hypotheses of the tree of life are reassuringly congruent with larger datasets. The limited ways the data could be analyzed emphasizes the need for advanced methods that utilize gene tree conflict to inform species relationships. To fully embrace the information provided by gene tree conflict, this thesis presents a new method of analysis that calculates the likelihood for a single relationship on the phylogenetic tree. By isolating a single relationship of interest, the other relationships are allowed to vary, which provides a means of incorporating the gene tree conflict into the analysis of the relationship. This method is used on the contentious relationship in the carnivorous clade and further validated using a vertebrate dataset. The results show that while the phylogenetic reconstruction supports one hypothesis, the data itself supports an alternate hypothesis. This presents a novel framework to conduct analyses that are focused on a single contentious relationship at a time. Overall, this thesis results in the development of novel methods for phylogenomic analyses to provide evolutionary insights into the hyperdiverse Caryophyllales.

Chapter I

Introduction

Background

Mutations provide the raw natural variation upon which natural selection may act. Over time, if they become fixed in a population, these substitutions provide a major source of information for us to reconstruct macro-evolutionary history across the tree of life. In 1965, Zuckerkandl and Pauling proposed the idea of using molecules, which carry the genetic information underlying the phenotypic traits we typically associate with adaptations, for inferring evolutionary history. This coincided with the English translation of Willi Hennigs "Phylogenetic systematics" and the introduction of a method for deducing branching sequences of a phylogeny (Camin and Sokal, 1965). The inference of phylogeny, which had previously only scattered the literature (Tillyard 1921, Haeckl 1865, Darwin 1859), grew into a major branch of evolutionary biology.

Early phylogenetic work utilized morphological characters to infer species relationships (Sokal, 1958; Wilson, 1965). This was largely due to the high cost of sequencing data. However, despite the limited data, models of molecular evolution were being actively developed. Setting the stage for the day molecular data, specifically nucleotide and amino acid, became affordable (Jukes and Cantor, 1969). The use of models allows scientists to incorporate saturation into the analysis (when a nucleotide has undergone multiple substitutions) and over time they have advanced and been put into a statistical framework. This allows inference of species relationships to more realistically incorporate the complexities associated with evolutionary change at the molecular level (Felsenstein, 1983). Despite the philosophical arguments and computational limitations, statistical phylogenetics won over the community and DNA began to become the standard information source for phylogenetics.

A new era for systematics

By the 1990's, DNA was the most popular data source for biologists seeking to understand evolutionary relationships and increased computational power led to modern statistical phylogenetics arising out of the methodological framework developed in the preceding decades. Researchers typically sequenced a small number of gene regions and when species relationships conflicted between studies or genes, this was often considered to be due to errors in methodology and thus the solution proposed was more advanced methods. As our knowledge of genomes expanded along with dataset size, it became clear that conflict could also be attributed to a multitude of biological reasons (Maddison, 1997). This theoretical development raised questions about what to do with data supporting different hypotheses. While methodology and theory continued to advance, technological advancements driven by genome sequencing ushered in a revolution in molecular data acquisition. As next-generation sequencing facilitated the gathering of enormous dataset, the issue of conflicting phylogenetic results became even more prevalent. Although these large datasets provide an unimaginable amount of evolutionary information, they also show how complex the process of speciation truly is.

The use of genomic data in phylogenetics (called phylogenomics), makes it impossible to ignore that a gene tree is not the same as a species tree. For example, phylogenomic analyses tend to always give perfect support when using traditional statistical support measures despite significant underlying conflict. The initial promise that genomes would give rise to a perfectly inferred tree of life was quickly hampered. For example, the most common measure used was the non-parametric bootstrap (Felsenstein, 1985). This method assumes that all sites in the sequence data examined are independently and identically distributed (i.i.d). Furthermore, the method assumes that all gene trees share the same topology, which in the case of phylogenomics is almost always violated. Thus the bootstrap, which is inarguably the most widely adopted support measure of phylogenetics, cannot detect poor support in phylogenomics. The inability to detect poor support has led to the development of a number of novel support measures designed to handle large datasets (Seo, 2008; Hoang, 2017; Pease et al. 2018). For these support measures, it is imperative that the different evolutionary histories among genes are accounted for since conflict may be biological and not systematic error. This has helped transform phylogenomics into a process-based field and brings new information about the evolutionary history that goes beyond species relationships.

Gene tree conflict emerges as a major aspect of phylogenetics

There are many biological processes that lead to genes having conflicting evolutionary histories (i.e., gene tree conflict). One of the most prevalent sources of gene tree conflict is incomplete lineage sorting (ILS), where the individual substitutions in genes do not have time to fix in the population of species. This is especially common in the relationships of species that have undergone rapid radiations. Another driver of gene tree conflict is hybridization, as seen in the case of humans sharing Neanderthal DNA (Fu et al. 2015) and Heliconus butterflies mimicry genes (Dasmahapatra et al. 2012). Furthermore, horizontal gene transfer may influence the relationship of genes, something seen for beneficial adaptations such as C4 photosynthesis (Christin et al. 2012). Methods specifically designed for phylogenomic data focus on ILS, often times the most prevalent form of conflict, however, these methods work on the assumption that the rate of mutation is the same for all genes and all individuals in the same branch of a species phylogeny (Mirarab et al. 2014). These assumptions are violated by natural selection and do not account for many of the biological processes that generate gene tree conflict.

Another major source of gene tree conflict is systematic error. For phylogenomic datasets, a major source of systematic error arises from improper orthology identification (Brown and Thomson, 2017). Homology is a fundamental question in evolutionary biology and, for phylogenomics, the field required the development of novel methods of homology detection (Yang and Smith, 2014; Yang et al. 2017; Walker et al. 2018). Genes may be duplicated, deleted, or modified from previous genes (Ohno et al. 1968). Proper phylogenetic analysis relies on shared evolutionary history and proper orthology detection. Many homology detection methods for phylogenomic datasets were fundamentally bioinformatic exercises without the input of phylogeny. However, by including phylogenies in these analyses, the results can be more accurate and greatly improved (Gabaldón, 2008).

This dissertation focuses on advancing phylogenomic methods, and using the data to infer evolutionary events underlying the data. Although, phylogenetics typically seeks to resolve species relationships, this dissertation seeks to embrace phylogenomic data as a means of understanding the evolutionary processes that have led to these relationships. The discovery that genes often do not share evolutionary histories, also brought to light a whole new use for the phylogenetic methods that have been perfected for the past 50 years. The focus for advancing methods involves three chapters, with three primary goals. **1)** Identify the issues associated with phylogenomic analyses. **2)** Develop an approach to homology detection capable of creating densely sample phylogenomic matrices and cutting the computational burden associated with homology detection in phylogenomics. **3)** Develop a method specifically designed to incorporate gene tree conflict into an analysis, with a focus on resolving a single contentious lineage, as opposed to all species relationships. While methodology is one aspect of these studies, another focus of this dissertation is on advancing our understanding of flowering plant evolution.

The methods developed in this thesis focused on improving our understanding of the clade Caryophyllales (Fig. 1-1), a hyperdiverse order of flowering plants. Comprising roughly 13,500 species, the Caryophyllales exhibits a cosmopolitan distribution, and has representatives on all seven continents of the earth (Hernández-Ledesma et al. 2015). The depauparate lineages that comprise the backbone (Cuénoud et al. 2002; Brockington et al. 2009; Hernández-Ledesma et al. 2015), mixed with the contentious relationships among the carnivorous plants makes Caryophyllales ideal for using phylogenomics to examine gene tree conflict. Throughout this process I also advanced our knowledge on the macroevolutionary patterns seen in the clade, by identifying high levels of polyploidy in the noncore Caryophyllales, and regions of uncertainty for major relationships in the clade.

Chapter summaries

In **chapter II,** I investigate sources of error in phylogenomic studies. I use the chloroplast genome "plastome" as a means of studying sources of gene tree conflict across the angiosperms broadly. The plastome contains 79 highly conserved genes across angiosperms. Typically, those genes should not conflict with each other, aside from in rare instances as the plastome is uniparentally inherited, and thus all genes are inherited as a single unit together.

This chapter demonstrates that the plastome contains a substantial amount of inferred gene tree conflict. When statistical support measures are considered, most of this conflict disappears, and is found to be the result of the gene being uninformative. This indicates much of the inferred conflict is likely systematic error. A further examination into this conflict shows some correlates that may lead to gene tree conflict, including alignment length and tree length (a proxy for information). Furthermore, misalignment of a single gene is shown to create enough signal to alter an inferred species tree. Overall, this helps provide some guidelines on assembling datasets in phylogenomics and provides evidence that not all genes should be included in an analysis.

In **chapter III**, I focus on the noncore Caryophyllales, a group of 2200 species with an array of charismatic carnivorous plants (e.g pitcher plants, venus fly trap, and sundews). Many species within this clade have long fascinated biologists and the general public alike. The pitcher plants contained in the genus *Nepenthes* were named after the potion drank by Helen of Troy to relieve her pain after the battle of Troy. Upon discovering such a fascinating plant, botanists named it after the potion believing that its existence was enough to cure sorrow. The Venus flytrap, is often anthropomorphized due to the trap resembling mouth, and its very existence called Linnaeus to question the Order of nature as willed by God. Finally, the Sundew inspired Darwin to write an entire volume on insectivorous plants, one in which while crafting his manuscript on the plant he famously lamented to his mentor Charles Lyell, that he cared more about sundews then about the origin of all species. This is often taken out of context to mean he cared significantly more about sundews than other species, however, the full letter unveils his frustration at writing a manuscript on the subject.

This clade, with its storied history, provided a perfect group to examine the performance of modern phylogenomic analysis, specifically through the use of transcriptome data. In this chapter I build upon the body of work performed by many other scientists, to find that the relationships commonly debated about in the literature are surrounded by high levels of gene tree conflict. Furthermore, I found evidence for at least seven paleopolyploidy events, with each family containing a unique one, aside from the monotypic family *Drosophyllaceae*. While performing more in-depth analyses into the commonly debated relationships, I found that, despite the use of 1237 genes for the species tree inference, the removal of a single taxa altered the species relationships. This demonstrates the sensitivity that still exists in even the most data rich analyses.

In **chapter IV,** I address the issue of sampling in large datasets. Although, these datasets provide a large amount of evolutionary information, the computational burden is a major limitation, especially for homology detection. Typically, researchers perform an all-by-all similarity search that scales quadratically. This chapter explores a faster method for detecting homology in a dataset of >300 genomes and transcriptomes.

To avoid doing the all-by-all blast procedure, I developed a method of hierarchical clustering, where previously inferred evolutionary relationships may be taken into account. Specifically, I developed a hierarchical clustering technique that allows for a divide-and-conquer approach to homology inference. I tested this method on more than 300 transcriptomes and genomes from the Caryophyllales.

Along with containing the non-core Caryophyllales, the Caryophyllales contain a hyperdiverse array of fascinating other plants. The charismatic family Cactaceae are known for their extreme radiations into desert climates. The group also contains economically important Amaranthaceae and Chenopodiaceae that include beets, quinoa, and spinach. Furthermore, this group provides a fascinating natural laboratory upon which scientists may study convergent evolution, with several independent convergences on the adaptations of C4 photosynthesis, cushion plant growth, and halophytism.

In addition to the methodological developments presented in this chapter, these analyses showed that many previously reported relationships were supported by phylogenomic data, and many of the most contentious relationships, had high levels (>50%) conflicting gene tree relationships. This was especially true in the family Cactaceae, known for its rapid radiation. This analysis also brought molecular evidence for a new family Agdestidaceae, previously proposed for its extreme morphological diversity this provided strong evidence for its circumscription. Several other taxonomic and systematic revisions have resulted from these analyses.

Overall, the contribution of hierarchical clustering proved a powerful means of combining what was previously an unfeasible amount of data. The biological results provide a framework for the focus of future studies in Caryophyllales. High levels of congruence with previous phylogenetic studies indicates that phylogenomics may be best used as a tool for understanding processes, and multi-gene phylogenies with traditional markers are a powerful tool for inferring species relationships.

In **chapter V,** I look to revisit two highly contentious relationships in plant and vertebrates. The plant relationship involves the non-core Caryophyllales and the vertebrate relationship involves the placement of alligators and caimans in relation to other vertebrates. In both cases, the concatenated maximum likelihood (ML) result was disregarded in favor of the relationship inferred from coalescent analyses. For both clades the coalscent results were more sensible biologically. For example, the coalescent result for the plants favors an *Ancistrocladus*+*Drosophyllum* clade. This reflects many shared morphological traits. The coalescent result for the vertebrate dataset alligators and caimans as sister to birds. This is supported by both alligators and birds sharing similar nesting behaviors, and other traits (e.g., they both chirp at birth). The conflicting datasets present complementary tests for new methods in examining conflict.

In this chapter, instead of developeing a new method for reconstructing entire topologies, I focused on examining a single relationship. This approach, the Maximum Gene Wise Edge (MGWE) approach, allows for relationships outside of the relationship of interest to vary between genes. I compared this to traditional approaches for examining conflict. The MGWE performed well when compared to the traditional means of inferring phylogenies, and showed support for the topology found using coalescent methods. This approach provides an early view into how phylogenomic data may be disentangled from traditional phylogenetic analyses to ask targeted questions.

Chapter VI provides brief concluding statements regarding the results of this work including a discussion on how phylogenomics has begun differentiating itself from phylogenetics and some thoughts on future directions for the field.

Concluding remarks of the Introduction

This dissertation seeks to expand our knowledge on the complex evolutionary history of the clade Caryophyllales through the development of novel methods and analyses. The approaches used here seek to take advantage of the wealth of data available through next generation sequencing. Throughout, a common theme of differentiating the genomic data from traditional phylogenetic analyses ties the chapters together, with an emphasize on the role gene tree conflict plays in inference. This is in part a result of discovering that phylogenomics is not only a field of inferring species relationships, but a field of inferring the processes that give rise to these inferred relationships.

Our use of phylogenetic trees for understanding how life on earth is related is remarkably similar to that of the views of someone who believes punctuated equilibrium to be the process of evolutionary change. All change is interpreted as speciation and nodes are instantaneous moments in time. In reality Charles Darwin's proposition of gradualism is likely closer to the truth, which leads species to be on a continuum. Phylogenomics allows us to investigate this continuum and as we become rich with data, we are beginning to embrace it through novel methods. This has led to another fascinating transformation of the field of systematics and makes our current day in age one of the greatest times to pursue how life on earth is interconnected through descent with modification. Although it would appear the concept of a species is disappearing with the genomic age, I'd like to end this introduction with a century old quote from "Problems of Genetics" by William Bateson.

"*But the experience of the practical breeder does, I think, on the whole, support the contention to which systematists have so steadily clung under all the assaults of evolutionary philosophers, that, though we cannot strictly define species, they yet have properties which varieties have not, and that the distinction is not merely a matter of degree*" –W. Bateson

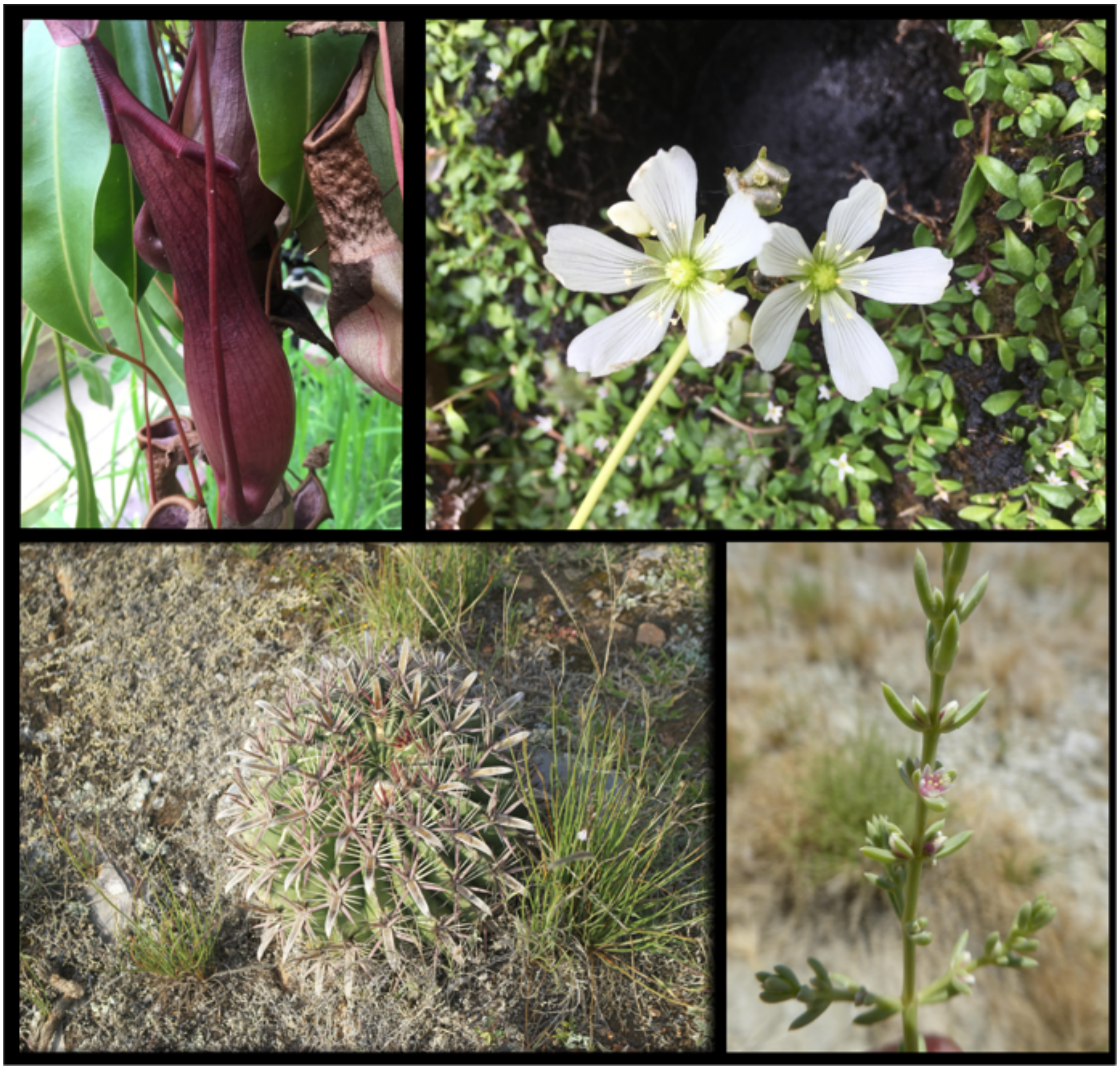


Figure 1-1 Example species of Caryophyllales

Clockwise from top left: The pitcher plant (*Nepenthes alata*), The venus fly trap (*Dionaea muscipula*), Cactus (*Ferocactus latispinus*), and the Polycnemoideae (*Nitrophila occidentalis*).

Chapter II

Characterizing gene tree conflict and systematic error in plastome-inferred phylogenies

Preamble: This chapter is our manuscript that is currently in review, the citation for this manuscript is: *JF Walker, GW Stull, N Walker-Hale, DA Larson and OM Vargas. Characterizing gene tree conflict and systematic error in plastome-inferred phylogenies.*

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 gene length is the strongest correlate to 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 systematic error is likely the root of most of the observed conflict, but further research on biological conflict within plastome is warranted given the documented cases of inter-plastome recombination. We suggest *rpoC2* as a useful marker for reconstructing angiosperm phylogeny, saving the effort and expense of assembling and analyzing entire plastomes.

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, 2011), and the resulting phylogenies have been used for countless other 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, 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, 2010; Ruhfel *et al.*, 2014; Stull *et al.*, 2015; Gitzendanner *et al*., 2018). 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.

Most gene tree conflict is 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 *et al*., 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 uniparentally inherited (usually maternally, with notable exceptions: e.g., McCauley *et al*., 2007) and undergoes a unique form of recombination that is not expected to result in conflicting gene histories within a single genome (Palmer, 1983; Bendich, 2004; Walker *et al.* 2015). However, instances of inter-plastome recombination have been documented in different plant groups (Sancho *et al*., 2018; Sullivan *et al*., 2017). Additionally, sharing of genes between the chloroplast and nuclear genomes remains another potential source of biological conflict (Martin *et al*. 1998; Martin 2003). Although biological conflict in the plastome generally seems rare, the true extent of intra-plastome conflict is poorly known given that the vast majority of studies assume no conflict as an operating principle.

Biological conflict aside, there remain significant potential sources of systematic conflict that have been poorly explored across the plastome (e.g., Burleigh & Mathews 2007a,b). 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 (Walker *et al*., 2014), as well as among individual genes (e.g., Shepherd *et al*., 2008). The rate of chloroplast evolution as a whole has been examined (and compared with the nuclear and mitochondrial genomes; Wolfe, 1987), and rate variation within the chloroplast—especially across the three major regions of the genome, i.e., the long single-copy (LSC) region, the short 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, 2007, 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 inferred from all three genomes (e.g., Soltis *et al*., 2011; Wickett *et al*., 2014; Zeng *et al*., 2014).

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, alignment method, taxon sampling, and data type (i.e., nucleotides vs. amino acids) at varying time scales across angiosperms. Our results show that the plastome—at all levels—contains notable gene tree conflict, with the number of conflicting genes at each node often comparable to the number of concordant genes; however, the majority of plastid genes are uninformative for most nodes when considering support. We reveal several sources of systematic error (e.g., poor alignment) contributing to conflict, but further work will be necessary to explore other potential causes of intraplastome conflict (e.g., inappropriate models, stochasticity, 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. Our results provide an important glimpse 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*; Appendix A). Our sampling scheme was designed to capture all major angiosperm lineages (e.g., Soltis *et al*., 2011), 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, from species-level relationships in *Diplostephium* (Asteraceae) to the ordinal-level relationships defining the backbone of the currently hypothesized angiosperm phylogeny. The data were divided into two sampling sets, one containing all 51 angiosperms (ALL), and another containing a subset (SUB) of ten of the angiosperm samples (defining the skeleton of angiosperm phylogeny); both datasets contained the two gymnosperm outgroups. The goal of the two sampling schemes was to determine if increased taxon sampling was helpful for alleviating gene tree conflict (given that improved taxon-sampling generally improves phylogenetic inference).

Data preparation, alignment, and phylogenetic inference

All scripts used and developed for this study 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 almost complete gene occupancy apart from instances of gene loss or pseudogenization. For all genes in the ALL and the SUB datasets, the amino acid and nucleotide data were aligned using Fast Statistical Alignment (FSA; 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 (Redelings, 2014), and does not rely upon a guide tree for sequence alignment, helping avoid downstream bias. The amino acid alignments were used to guide another nucleotide alignment (which we refer to as the "codon alignment" or "codon-guided alignment") using the phyx program pxaa2cdn (Brown *et al.* 2017). 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/codon aligned 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 program pxcat. For the nucleotide, codon, and amino acid data for the ALL and the SUB datasets, we inferred plastome phylogenies using the GTR+G and the PROTGAMMAAUTO models as implemented in RAxML. To complement the model inference performed by RAxML from the AUTO feature, we also used IQ-TREE's (Nguyen *et al.* 2014) built-in model selection process (Kalyaanamoorthy *et al*. 2017) on the partitioned data.

Determination of outlier genes

Using the partitioned supermatrices for the three SUB datasets, we performed a site-specific log-likelihood (SSLL) analysis, with the model settings mentioned above. The comparison was made between the codon-inferred topology and the 'true topology' (TT) of angiosperms (i.e., a tree summarizing our current understanding of angiosperm phylogeny from the literature: e.g., Soltis et al. 2011; Moore et al., 2007, 2010; Wickett et al., 2014); both the nucleotide and amino acid data sets inferred the TT, so only the codon-inferred tree was compared with the TT to identify outliers potentially driving the erroneous codon topology. Analyses imposing the codon topology and TT were run for all datasets. We then summed the SSLLs for each gene to obtain the gene-wise log-likelihoods. This was also performed for the ALL datasets, where the tree inferred from each plastome supermatrix was compared to that of the TT. This was done because, in the case of the ALL datasets, the topology inferred from each plastome (i.e., each set of concatenated genes) was unique compared to the TT and the other ALL datasets. Outlier genes evident from the gene-wise log-likelihood calculations were then examined for anomalies in topology, tree length, and root-to-tip variance (see **Assements of Individual Plastid Genes** below), all of which are useful for highlighting the possibility of systematic error.

Analysis of conflict

All gene trees were rooted on the outgroups using the phyx program pxrr (Brown *et al.* 2017), in a ranked fashion in the order *Podocarpus lambertii* then *Ginkgo biloba*. Conflict in the data was identified using the bipartition method as implemented in phypartsv.0.0.1 (Smith *et al.* 2015), with the gene trees from each data set (amino acid, codon, and nucleotide) mapped against the TT described earlier. The concordance analyses were performed using both a support cutoff (at 70% bootstrap support, i.e., moderate support) and no support cutoff. When the support cutoff is used, any gene tree node with under 70 bootstrap support is regarded as uninformative for the 'species tree' node in question; when no support cutoff is used, the relationship in the gene tree factors into the concordance/conflict of the species tree regardless of the support value. In both cases, a gene is considered uninformative if a taxon relevant to a particular node/relationship is missing from the gene dataset. The analyses in which support was taken in to account were visualized using the script phypartspiecharts.py (github.com/mossmatters/MJPythonNotebooks/blob/master/phypartspiecharts.py). Conflict was also analyzed based on the estimated time of divergence for each node/clade (see Fig. 2-1 for time bins). The nodes were binned based on their inferred ages (ages > 20 Ma from Magallón *et al*., 2015; ages < 20 Ma from Vargas *et al*., 2017 and Roquet *et al*., 2009) into one of five categories, each representing a time interval of 30 Ma (starting roughly with the origin of angiosperms, at 150 Mya, to the present). 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.

Assessments of Individual Plastid Genes

Using the phyx program pxlstr (Brown *et al.* 2017), we calculated summary statistics for each gene alignment and corresponding gene tree: number of included species, 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 different measures of a gene's information content. Levels of concordance of each gene tree with the TT were then assessed by tabulating the number of nodes concordant between the gene tree and TT. The number of concordant nodes (in Fig. 2-2 this is treated as a proportion of total nodes available to support and in Fig. 2-3 this is based on total nodes) was used as a measure of the gene's ability to accurately reconstruct angiosperm phylogeny.

We examined the relationships between gene tree concordance and alignment length, tree length, and root-to-tip variance using logistic regression, considering each node as a trial and considering trials in aggregate across genes. Because large predictor values of some genes rendered them highly influential, we also conducted analyses excluding some genes. We tested specific hypotheses of the relationship between alignment length and informativeness using regressions of alignment length and tree length only. Details of these analyses are given in the supplementary materials (Appendix A).

We also performed saturation analyses on all the chloroplast genes to determine if they were capable of inferring deep divergence times (see Appendix A).

Comparison of genomic regions

We assessed the utility of the three major plastome regions—the Long Single Copy (LSC) region, Short Single Copy (SSC), and the Inverted Repeat (IR) region—for reconstructing angiosperm phylogeny in two ways. First, we constructed ML phylogenies (as described above for each of the 'plastome' analyses) for each genomic region using the concatenated set of genes comprising each region; with the resulting trees, we then calculated (for each genomic region) the number of nodes concordant with the TT of angiosperms. 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 (Appendix A); this permits a qualitative visual assessment of the general concordance levels of each genomic region at each time slice.

Results

Patterns of conflicting chloroplast signal

The sampling for this experiment allowed us to examine conflict at multiple evolutionary scales, from species-level relationships to ordinal-level relationships. Our main sampling ('ALL') included 51 angiosperms and two gymnosperm outgroups; our reduced datasets ('SUB') included 10 angiosperms and two gymnosperm outgroups. For both sampling levels, we examined amino acids, codon-aligned nucleotides, and non-codon aligned nucleotides. First, we compared the topology inferred from each concatenated set of protein-coding genes (hereafter referred to as 'plastomes') to the 'true topology' (TT) of angiosperms (i.e., a tree summarizing our current understanding of angiosperm phylogeny from the literature: e.g., Soltis *et al*. 2011; Moore *et al*., 2007, 2010; Wickett *et al*., 2014). Our plastome trees for both the ALL and the

SUB datasets were highly concordant with the TT (Fig. 2-1) with the exception of the concatenated codon alignments (see the section **Analysis of outlier genes** below). Without considering support, the gene trees showed notable levels of conflict across the different analyses (Figs. 2-1–2-3). When considering support, the majority of the plastid genes were uninformative for practically all nodes in the phylogeny (Fig. 2-1; Appendix A); 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 amount of gene tree conflict and evolutionary scale (i.e., conflict was relatively evenly distributed across shallow and deeper nodes/time scales; Fig. 2-1, 2-2). Although the greatest degree of gene tree concordance with the TT appeared in the nodes with inferred ages between 90–61 mya (ages based on Magallón et al., 2015), these nodes typically still contained at least 50% uninformative gene trees (Fig. 2-1). Instead, analysis type (whether examining amino acids, codon-aligned nucleotides, or non-codon-aligned nucleotides) had a much greater impact on the prevalence of conflict (Figs. 2-2, 2-3), with the amino acid dataset generally showing higher levels of gene tree conflict. When factoring in support (BS 70 cutoff), the amino acid data set showed even less concordance with the TT (as more genes were considered uninformative due to low BS support); the integration of support also decreased concordance of the nucleotide data sets with TT, but proportionally less. See Appendix A for conflict analyses showing the amino acid, codon, and nucleotide (ALL) gene trees mapped onto the TT (the codon results from Appendix A are also shown in Fig. 2-1). Although the topology inferred from the concatenated codon-aligned genes was vastly different from the TT (due to the presence of two outlier genes; discussed more below; Fig. 2-4), the codon-aligned genes and nucleotide alignments showed roughly equivalent levels of concordance with the TT (Fig. 2-3), regardless of whether or not bootstrap support was considered.

To see 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 our data—rather than just a single model for the entire concatenated set of genes. 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 80 genes based upon the IQ-TREE model test (Appendix A).

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). We used logistic multiple regression to test relationships between gene performance and characteristics (Appendix A). We found that alignment length had a significant positive multiplicative relationship with odds of concordance across all datasets (Fig. 2-2, Appendix A). Tree length and root-to-tip variance had significant positive and negative multiplicative relationships, respectively, in amino acid and nucleotide datasets but not codon datasets when allowing for overdispersion (Appendix A). Notably, excluding highly influential observations with outlying predictor values rendered all three predictors significant but did not affect the direction of most relationships, except for tree length which had a significant positive multiplicative effect in codon datasets (Appendix A). 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. 2-2, Appendix A).

Genomic patterns of concordance/conflict

In terms of number of nodes concordant with the TT, the LSC and SSC regions were roughly comparable, with both outperforming the IR in terms of number of nodes concordant with the TT, whether considering bootstrap or not (Appendix A). 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 alignment lengths of the LSC region, SSC region, and IR were 73422 bp, 10395 bp, and 19314 bp respectively. The genome diagram of concordance (Appendix A) 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) with the exception of 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 TT (Fig. 2-2, Appendix A); in general, it performed at least as well as the 79 concatenated genes in reconstructing the TT. 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, performing relatively poorly (Fig. 2-2). The *matK* alignment is ~250 bp longer than the *rbcL* alignment; the best performing gene, *rpoC2* (alignment length 4660 bp), is one of the longest plastid genes. However, the notably long region *ycf1*—which encodes for ca. ~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. 4660 bp for *rpoC2*). In several cases it performed toward the top; however, it was never the top-performing gene in terms of number of concordant nodes

(Appendix A). Notably, despite the high levels of observed conflict overall, we found that every node of the TT was supported by at least one gene. Thus, to varying degrees, all relationships of the TT are found within the plastome gene tree set.

Analysis of outlier genes

For the ALL datasets, plastome phylogenies inferred from the amino acid data and nucleotide data (non-codon aligned) were broadly concordant with the TT, with only minor differences in consistently problematic parts of the angiosperm tree (Figs. 2-1, Appendix A)—e.g., relationships amongst lamiids orders. For the SUB datasets, both the nucleotide data (non-codon aligned) and amino acid data inferred a topology identical to the TT. However, the codon-guided nucleotide alignment inferred a plastome phylogeny that differed substantially from the TT, for both the SUB (Fig. 2-4) and ALL (Appendix A) datasets. A gene-wise log-likelihood (GWL) (Shen *et al.* 2017) comparison between the TT and codon-alignment tree showed that only four genes supported the topology inferred by the codon-aligned supermatrix (Fig. 2-4); the rest supported the TT.

While most genes exhibited minimal signal, *ndhD* and *rpl2* showed vast likelihood differences in favor of the topology inferred from the codon-aligned nucleotide alignments (Fig. 2-4; Appendix A). Exceptionally, *ndhD* had a >4000 likelihood difference toward the codon-inferred topology for the SUB dataset and >17000 for the ALL dataset. These extreme likelihood differences indicate that these two genes (which were 1563 and 954 bp in alignment length, for *ndhD* and *rpl2*, respectively) were driving the entire concatenated codon-based supermatrix toward a radically incongruent topology of angiosperms. The tree root-to-tip variance (251.559, 60.3827) and tree length (35.2016, 16.0043) of *ndhD* and *rpl2*, respectively, were notably high compared to the remaining plastid genes (Appendix A), highlighting potential problems with these two regions. Further examination of the genes revealed poor alignments stemming from errors in GenBank submissions (discussed further below), indicating that these aberrant phylogenies are an artifact of systematic error. We also examined the nucleotide and amino acid alignments and found that one gene from the nucleotide set and five genes from the amino acid set supported the codon-inferred topology, even though overall each dataset supported the TT and they had no obvious sources of systematic error.

Discussion

Our use of a reference phylogeny (or 'true topology', TT), based upon numerous previous studies (e.g., Soltis *et al*., 2001, 2011; Moore *et al*., 2007, 2010; Wickett *et al*., 2014), provided us a benchmark against which the gene trees and 'plastome' trees (inferred by the three supermatrices: amino-acid, nucleotide, and codon-aligned nucleotide) could be compared. The TT is a synthesis of results from all three plant genomes (nuclear, chloroplast, and mitochondrial) and is treated here as a hypothetical species tree; this allowed us to better evaluate conflict/concordance of the 'plastome' (i.e., here, the 79 concatenated protein-coding genes) as well as the individual plastid gene trees with the angiosperm 'species tree'. Our expectation, based upon the chloroplast's mode of inheritance, was that all genes in the plastome should have the same history. Therefore, we expected that all plastid genes should show similar patterns of conflict when compared with non-plastid inferred phylogenies. Furthermore, conflicting relationships between the 'plastome' and the TT should be identical across data types (amino-acid, nucleotide, codon-aligned nucleotide). However, our results, discussed below, frequently conflict with this null model.

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 TT (Fig. 2-1; Appendix A). While the codon-aligned 'plastome' tree was initially drastically different than the TT (for both the ALL and SUB datasets), we discovered this to be a result of alignment error in two genes (discussed further below). For the amino acid and non-codon aligned nucleotide datasets, the ALL trees were highly similar (but not identical to) the TT, while the SUB trees were identical to the TT. The apparent increased accuracy of the SUB trees is almost certainly due to their highly skeletal sampling of angiosperm phylogeny lacking multiple recalcitrant nodes represented in the ALL trees.

The genes within the chloroplast genome are largely uninformative for most nodes of the phylogeny—however, a number of genes exhibited well supported conflict (Fig. 2-1). In general, there appears to be no relationship between evolutionary scale and amount of gene tree conflict: conflict generally does not appear confined to particular regions of the tree. Instead, the extent of conflict/concordance had a stronger relationship with data/analysis type (whether examining amino acids, codon-aligned nucleotides, or non-codon-aligned nucleotides). The amino-acid dataset showed the highest levels of gene tree conflict (Figs. 2-3, Appendix A), and both nucleotide datasets had about half the amount of gene tree conflict found in the amino acid data (Figs. 2-3, Appendix A). However, as noted above, several outlier genes in the codon-aligned dataset initially resulted in 'plastome' topology that conflicted strongly with the TT (discussed below). With the exception of outlier genes (discussed below), 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. 2-1, Appendix A). We suggest that the possibility of biological conflict deserves further exploration, especially given the documented instances of inter-plastome recombination (Sancho *et al*., 2018; Sullivan *et al*., 2017) and chloroplast-nuclear genomic exchange (Martin *et al*. 1998; Martin 2003).

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*), longer alignments should result in better-informed models, aided by both parsimony-informative and -uninformative characters (Yang, 1998). Inherent differences in amino acid and nucleotide models might also explain differences in performance. The nucleotide data was run using the GTR model, where the substitution rates between bases is individually estimated; however, 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 (Appendix A). This might be the result of the different methods implemented in RAxML vs. IQ-TREE for model testing, the different available models, 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 TT for the ALL sampling, 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) 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, 2018).

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; Appendix A), indicating that nucleotides are the most informative molecule 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.

Impact and detection of outlier genes in phylogenetic inference

Although the codon-aligned genes individually were more concordant with the TT than the other data types (amino acid and non-codon-aligned), the concatenated set of codon-aligned genes resulted in a highly incongruent phylogeny of angiosperms (compared to the TT). By examining the GWL of the codon-aligned genes—a measure of the influence of individual genes toward the selection of alternative topologies (i.e., here, the TT and the inferred concatenated codon-aligned tree)—we found that a vast proportion of the signal supporting the codon-aligned tree was derived from two genes, *ndhD* and *rpl2*, with only two additional genes providing support (albeit minimal) for that topology (Fig. 2-4). Closer examination of these outlier genes (*ndhD* and *rpl2*) revealed that this result stemmed from alignment error, due to errors in the GenBank submission: mismatches between the amino acids and nucleotides in the GenBank submissions led to an incorrect frame shift in the alignments. This underscores how minor systematic errors in only a few genes can have severe consequences for phylogenetic inference. In this case, poor alignment of only two genes resulted in a strong erroneous signal that was able to swamp the weaker signal present in the remaining 77 genes.

Interestingly we found that a number of genes from the nucleotide and amino acid alignments supported the (initially flawed) codon-inferred topology, albeit with minimal signal. The topology is semi-random for those alignments and thus the fact that they support the ‘incorrect’ topology likely speaks to their poor abilities as phylogenetic markers. Overall, our results present further evidence that a chloroplast analysis including all genes may not be favorable as some produce seemingly random signal and many genes are uninformative.

Several previous studies have highlighted the significance of outlier genes in phylogenomic analyses (Shen *et al.* 2017; Brown *et al.* 2017; Walker *et al.* 2018). These studies and our current results highlight the importance of screening phylogenomic datasets for outlier genes that might negatively influence analyses. Simple means of screening datasets include measuring the tree length (a measure of evolutionary rate) and root-to-tip rate variance (a measure of among-lineage rate variation) of all gene trees. Outlier genes in this case were shown to have extreme values relative to the rest of the genes (Appendix A). The topologies of extreme gene trees can then be examined for anomalies, and alignments and other preceding steps can be inspected for potential errors. In the present study, because the major relationships of angiosperms are well known, the negative impact of these outliers was easily detected. However, in study systems where the phylogeny is not well-known *a priori*, outlier genes could result in the inference of an incorrect phylogeny.

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, 2007, 2014) highlighted non-coding DNA regions useful for shallow evolutionary studies, while Graham and 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 'true' phylogeny of angiosperms (TT). 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 TT—however, *matK* performs considerably better (i.e., it generally has more nodes concordant with the TT; Appendix A**)**. This is likely due to a strong positive correlation between alignment length and number of concordant nodes, as noted above (Fig. 2-2); *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.*, 2008; Neubig and Abbot, 2010; Thomson *et al.*, in press) and barcoding (Dong et al. 2015), and here we find that it generally performs above average. The alignment of *ycf1* is abnormally long, 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* (Appendix A). 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 (Appendix A), and its alignment length (4660 bp) is ~1/5th the length of *ycf1*, easing the computational burden of using alignment tools such as FSA (which would struggle with a region as long as ycf1), as well as divergence dating and tree-building programs such as BEAST (Redelings and Suchard 2006; Drummond and 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 (Appendix A). When support is considered (Tables S2–S4), *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 TT). The utility of *rpoC2* likely stems from its notable length, resulting in a wealth of useful phylogenetic information. 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). This characteristic makes it particularly useful for comparative studies requiring accurate branch length estimates. 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 include non-coding regions in our study, which would likely outperform coding regions at shallow phylogenetic levels.

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. However, our results suggest that the coding sequences of the IR alone perform poorly compared to the LSC and SSC coding regions for reconstructing angiosperm phylogeny (Appendix A). However, 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. However, our saturation analyses (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, 2007, 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 (Appendix A), 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 can be a time-consuming process (Lanfear *et al.*, 2012; Kainer and Lanfear 2015), and which is beyond the scope of this study—might generally improve plastome-inferred phylogenies. Our results show that plastome datasets should be analyzed carefully, and that simple analyses of concatenated genes can produce poor or misleading results. More research on modeling molecular evolution across the chloroplast is needed.

Implications of plastid conflict for phylogenomic studies

Systematic error can be prevalent in phylogenomic datasets, as demonstrated here, with consequences for both phylogeny inference and downstream biological interpretation. In terms of inference, just a few outlier genes (resulting from misalignment, for example) can drive even large data sets toward the incorrect tree. This is because outliers often have large amounts of influence on inferred species relationships, as has been previously demonstrated (Shen *et al.* 2017; Brown & Thomson 2017; Walker *et al.* 2018). Also, when conflict is observed in phylogenomic studies, it should not be immediately assumed to be biological. Here we found clear examples of systematic error leading to conflict, with this error evident in the tree length and root-to-tip variance of outlier genes. The conflict we observed in the chloroplast—and the conflict observed in vertebrate mitochondrial genomes in a previous study (Richards *et al.*

2018)—suggests that systematic error might be an underappreciated source of conflict in phylogenetic studies. Chloroplasts—and organellar genomes in general—should harbor minimal amounts of biological conflict within them given their mostly uniparental inheritance (Birky *et al.*, 1995). Thus, they can serve as useful controls for examining the extent and sources of systematic conflict in phylogenomic analyses (e.g., Richards *et al.*, 2018, focusing on mitochondrial genomes). Here, we found the majority of plastid genes to be uninformative, but we also observed notable instances of well-supported conflict across multiple nodes of angiosperm phylogeny. Considerable gene tree conflict has also been documented in numerous phylogenomic studies using nuclear data (e.g., Rokas et al. 2003; Smith et al. 2015). The chloroplast, mitochondria, and nuclear genomes have very different modes of inheritance and rates of evolution (on average), which intuitively results in differences in the extent and types of gene tree conflict. However, the presence of intraplastome conflict implies that systematic error (stemming uninformative genes, alignment error, poor modeling, or other issues) might be a significant source of conflict in nuclear datasets as well. Furthermore, the finding that length plays a significant role in conflict, and that most genes are uninformative, should lead researchers to examine phylogenomic datasets with similar tools to those implemented here to ensure reliable results.

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. JFW was supported by a Rackham Predoctoral Fellowship. G.W.S. was supported by a NSF Postdoctoral Fellowship (NSF DBI grant 1612032).

O.M.V was supported by NSF FESD 1338694 and DEB 1240869. D.A.L. was supported by NSF DEB grant 1458466.

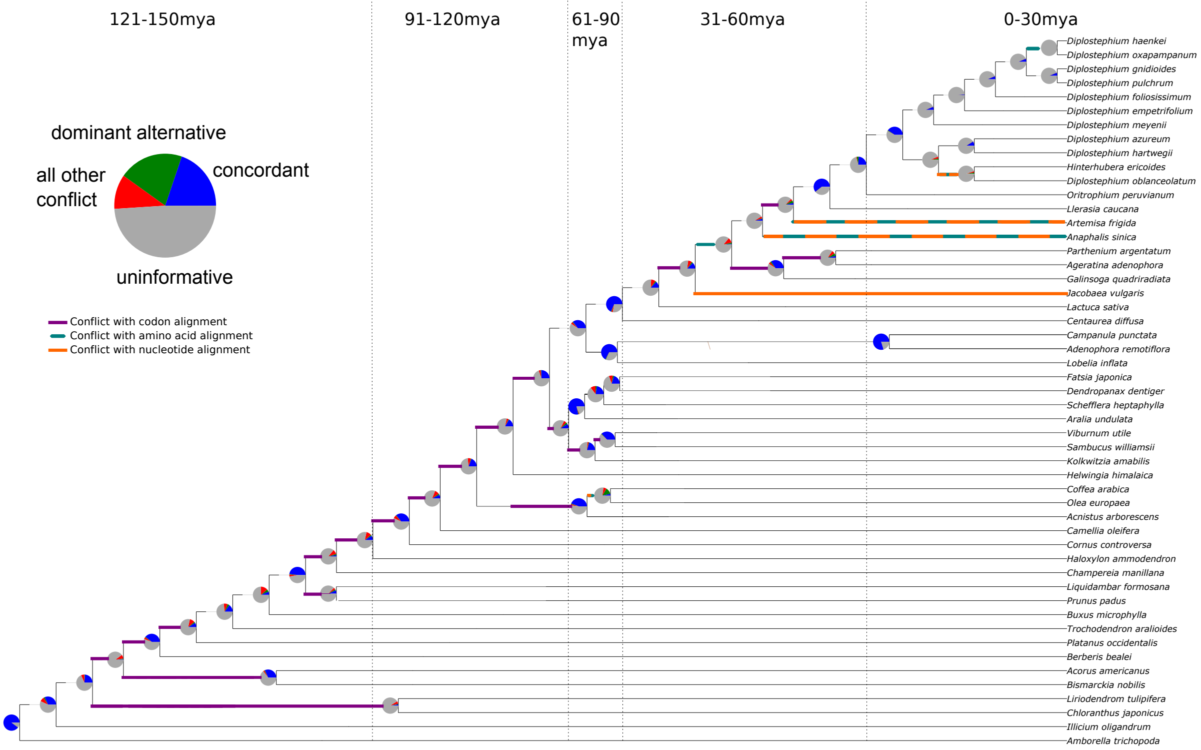


Figure 2-1 Summary of chloroplast conflict against the reference phylogeny of angiosperms

Purple, green, and orange lines indicate where the codon-, amino acid-, and nucleotide-inferred plastome trees conflict with the reference phylogeny. Pie charts depict the amount of gene tree conflict observed in the codon-based analyses, with the blue, red, green, 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 Magallon *et al.* 2015 and Vargas *et al.* 2017) used to bin nodes for examinations of conflict at different levels of divergence.

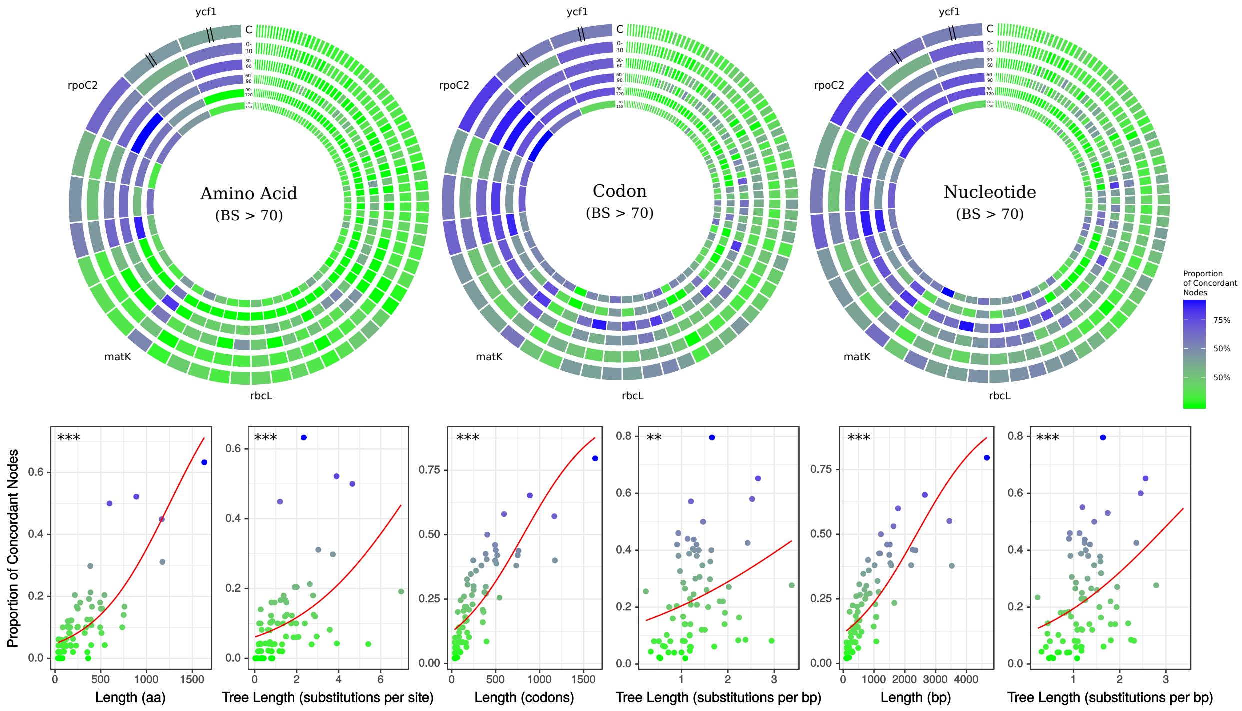


Figure 2-2. Gene tree concordance/conflict at varying time scales

Each diagram represents a different molecule type and shows the proportion of concordance each gene exhibits at the five time slices shown in Fig. 2-1: (1) 150–120 mya, (2) 120–90 mya, (3) 90–60 mya, (4) 60–30 mya and (5) 30–0 mya. The individual genes are scaled by length of alignment; however, *ycf1* and *ycf2* are cut to approximately the length of *rpoC2* due to their abnormally long alignments. The plots along the bottom show the relationships between gene concordance levels and various attributes of the genes (e.g., alignment length, tree length).

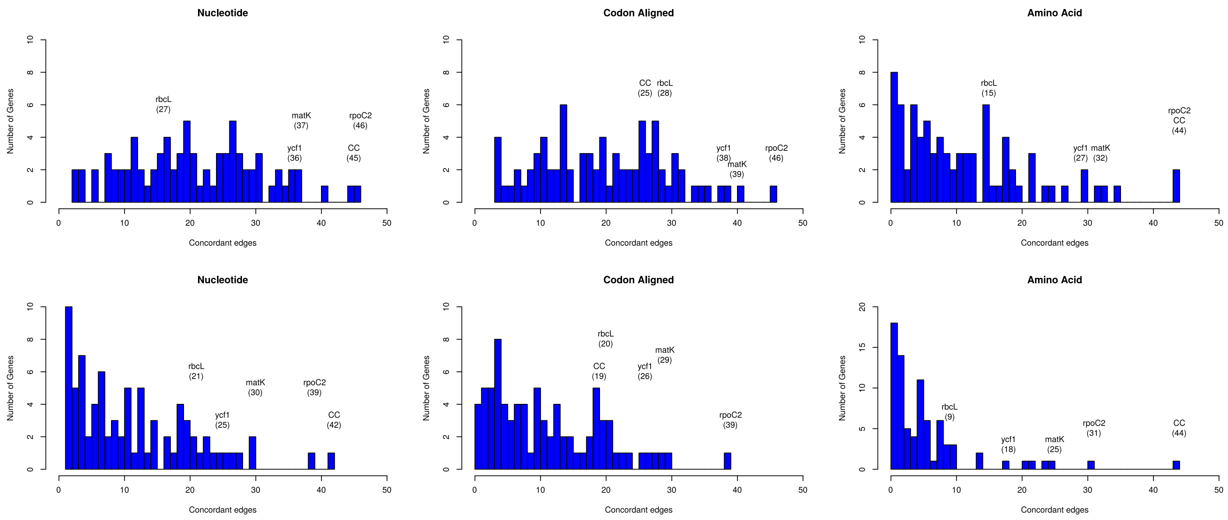


Figure 2-3 Histograms depicting number of concordant edges each gene tree contains compared to the reference phylogeny (i.e., the 'true tree', TT)

Each molecule type is plotted twice (once integrating BS support, shown on the bottom row; another time not integrating BS support, shown on the top row), resulting in six total histograms. The x-axes place each gene based on the number of concordant nodes it shares with the TT; the y-axes show the number of genes with different counts of concordant nodes. 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 TT.

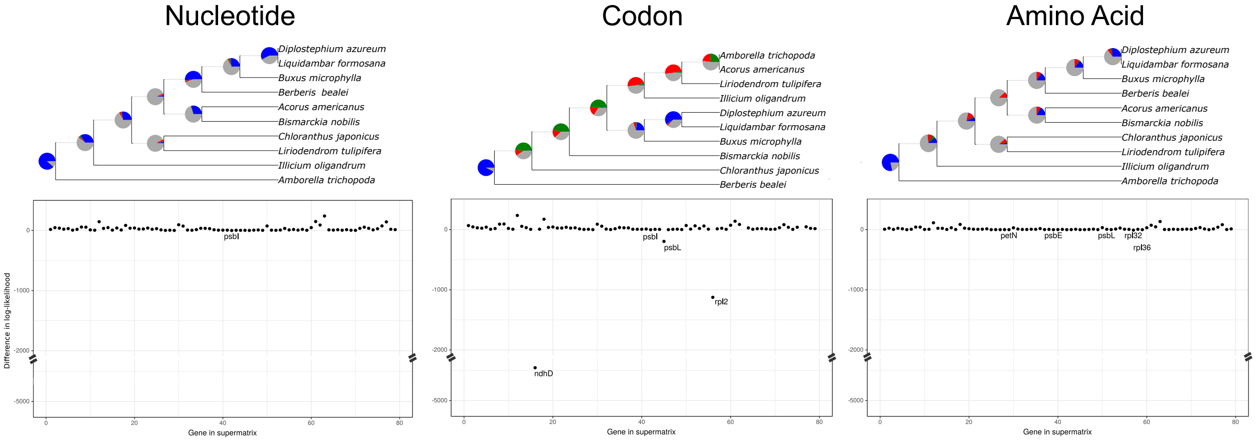


Figure 2-4 Depiction of the disproportion influence of outlier genes (due to misalignment) on topology resulting from analyses of the SUB datasets

The phylogenies are those inferred from each molecule type from the SUB datasets. The pie charts depict 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. Plots below the phylogenies are the difference in gene-wise likelihoods between the reference angiosperm phylogeny (positive values) and the codon-inferred phylogeny (negative values); the nucleotide and amino acid data yielded topologies identical to the reference phylogeny (i.e., the ‘true tree’, TT). Genes supporting the codon-inferred topology are labeled and breaks from ~(-2000, -4400) are place on the y-axis.

Chapter III

Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales

Preamble: This chapter is our manuscript that is currently in review, the citation for this manuscript is: *JF Walker, Y Yang, MJ Moore, J Mikenas, SF Brockington, A Timoneda and SA Smith*. Widespread paleopolyploidy, gene tree conflict and recalcitrant relationships among the carnivorous Caryophyllales (2017). *American Journal of Botany*, 104:6.

Abstract

The carnivorous members of the large, hyperdiverse Caryophyllales (e.g. Venus flytrap, sundews and *Nepenthes* pitcher plants) represent perhaps the oldest and most diverse lineage of carnivorous plants. However, despite numerous studies seeking to elucidate their evolutionary relationships, the early-diverging relationships remain unresolved. To explore the utility of phylogenomic data sets for resolving relationships among the carnivorous Caryophyllales, we sequenced ten transcriptomes, including all the carnivorous genera except those in the rare West African liana family (Dioncophyllaceae). We used a variety of methods to infer the species tree, examine gene tree conflict and infer paleopolyploidy events. Phylogenomic analyses support the monophyly of the carnivorous Caryophyllales, with a crown age of 68-83 mya. In contrast to previous analyses we recover the remaining non-core Caryophyllales as non-monophyletic, although the node supporting this relationship contains a significant amount gene tree discordance. We present evidence that the clade contains at least seven independent paleopolyploidy events, previously debated nodes from the literature have high levels of gene tree conflict, and taxon sampling influences topology even in a phylogenomic data set, regardless of use of coalescent or supermatrix methods. Our data demonstrate the importance of carefully considering gene tree conflict and taxon sampling in phylogenomic analyses. Moreover, they provide a remarkable example of the propensity for paleopolyploidy in angiosperms, with at least seven such events in a clade of less than 2500 species.

Introduction

Carnivory in plants has long fascinated both the general public and evolutionary biologists. Charles Darwin himself dedicated an entire volume to carnivorous species in his *Insectivorous Plants* (Darwin, 1875). The wide array of traps that are used to catch insects and other prey items make carnivorous plants some of the most morphologically diverse plants on Earth (Ellison and Gotelli, 2001; Heubl et al., 2006). These plants are able to occupy nutrient poor soils that would otherwise be unsuitable for plant life by obtaining nutrients unavailable in the soil through the digestion of animals.

Across angiosperms, carnivory is hypothesized to have independently evolved at least nine times (Givnish, 2015). One of these events is thought to have occurred relatively early on (~83 mya) in the non-core Caryophyllales (Magallón et al., 2015), giving rise to a “carnivorous clade” consisting of the fully carnivorous families Droseraceae, Drosophyllaceae, and Nepenthaceae, the small non-carnivorous African family Ancistrocladaceae, and the rare west African family Dioncophyllaceae, which includes the unusual carnivorous liana *Triphyophyllum peltatum* and two other monotypic, non-carnivorous genera (*Dioncophyllum* and *Habropetalum*)

(Albert et al., 1992; Meimberg et al., 2000; Brockington et al., 2009; Soltis et al., 2011; Hernández-Ledesma et al., 2015). The carnivorous clade of Caryophyllales comprises approximately 250 of the estimated 600 species of carnivorous angiosperms (Heubl et al., 2006; Ellison and Gotelli, 2009) and includes a diverse assemblage of trap-plants and pitcher plants that occupy a wide range of ecosystems, from the fully aquatic *Aldrovanda vesiculosa* to desert species of *Drosera* to the rainforest liana *Triphyophyllum*. Moreover, carnivory also appears to have been lost 1-3 times (Heubl et al., 2006) within the carnivorous clade, including in the ancestor of the 16 species of Ancistrocladaceae (Taylor et al., 2005) as well as in the ancestors of *Dioncophyllum* and *Habropetalum* in Dioncophyllaceae (Meimberg et al., 2000).

Despite broad appeal and interest, the evolutionary relationships in the non-core Caryophyllales remain ambiguous, with studies seeking to resolve these relationships often resulting in individually well supported but mutually conflicting topologies (Meimberg et al., 2000; Cameron et al., 2002; Brockington et al., 2009; Hernández-Ledesma et al., 2015). Much of this conflict involves the earliest branch in the non-core carnivorous clade, with studies finding Nepenthaceae as sister to the remaining lineages (Hernández-Ledesma et al., 2015), others finding Droseraceae as sister to the rest of the group (Meimberg et al., 2000), and yet others finding Droseraceae to be sister to the Nepenthaceae (Brockington et al., 2009). The strong support for conflicting topologies from different studies may be explained by the reliance on one or a few genes leading to systematic error (Maddison, 1997; Rokas et al., 2003). This type of error can arise from a variety of sources, including, but not limited to, incomplete lineage sorting, horizontal gene transfer, hybridization and hidden paralogy (Galtier and Daubin, 2008). Untangling these processes has proven to be a challenge and adds a strong level of complexity to phylogenomic analyses (Smith et al., 2015).

Transcriptomes have proven to be a powerful source of data for understanding this complexity, and have helped provide insight into the evolutionary history of non-model species (Dunn et al., 2008; Cannon et al., 2015; Yang et al., 2015). The thousands of genes typically sequenced in a transcriptome provide a means of identifying gene duplications and paleopolyploidy events (Cannon et al., 2015; Yang et al., 2015; Barker et al., 2016), which may clarify whether such events have been major drivers of evolutionary novelty (Ohno et al., 1968; Soltis et al., 2014). Moreover, analyses of gene tree concordance and conflict allows for a better understanding of the formation of species relationships and the complexity that arises in genomes as a result of speciation (Pease et al., 2016).

In this study, we conduct the first phylogenomic analysis focused on the non-core Caryophyllales, with sampling that covers all genera of carnivorous Caryophyllales except the poorly studied and rare lianas in the family Dioncophyllaceae of West Africa. We use large datasets to help resolve evolutionary relationships and explore gene tree discordance and its possible causes, as well as its consequences for phylogenetics among the carnivorous Caryophyllales. We find that, even with phylotranscriptomic data, many of the complications observed earlier in targeted sequencing studies (e.g. taxon sampling, gene tree conflict) are still present. However, we show how transcriptome data provide important insights into the reasons for these complications. Furthermore, we use transcriptome data to help provide information on the prevalence of paleopolyploidy in this ecologically and morphologically diverse clade and explore the molecular evolution of the group.

Materials and Methods

Data Availability

Raw reads for the ten newly generated transcriptomes were deposited in the NCBI Sequence Read Archive Appendix B (see the Supplementary Data with this article) (Bioproject: PRJNA350559). Assembled sequences, data files, programs, alignments and trees are available from Dryad.

Taxon Sampling, Tissue Collection, Sequencing and Data Assembly

The tissue collection, RNA extraction, library preparation, and quality control were carried out using a previously developed workflow (Yang et al., 2017). Transcriptomes of eight non-core Caryophyllales families representing nearly all of the major lineages of non-core Caryophyllales were included in this study (Appendix B). The transcriptomes of *Dionaea muscipula, Aldrovanda vesiculosa, Nepenthes ampullaria* and *Reaumuria trigyna* were downloaded from the NCBI Sequence Read Archive [accessions SRX1376794, SRR1979677, (SRR2666506, SRR2866512 and SRR2866533 combined) and (SRX105466 & SRX099851 combined) respectively] (Dang et al., 2013; Brockington et al., 2015; Bemm et al., 2016; Wan Zakaria et al., 2016). The assembly used for *Frankenia laevis* was the same as in Yang et. al. (2015) and can be found in Dryad (http://dx.doi.org/10.5061/dryad.33m48). The genomes of *Beta vulgaris* (RefBeet-1.2) and *Spinacia oleracea* were downloaded from The *Beta vulgaris* Resource (http://bvseq.molgen.mpg.de/Genome/Download/index.shtml; accessed Jul 10, 2015) (Dohm et al., 2014). We generated ten new transcriptomes for this study from fresh tissue collected from *Drosera binata*, *Nepenthes alata, Ancistrocladus robertsoniorum, Plumbago auriculata, Ruprechtia salicifolia* and *Drosophyllum lusitanicum*. The *D. binata* and *N. alata* data were also collected from trap tissue at three different developmental stages (Appendix B). The plant tissues were flash frozen in liquid nitrogen and stored at -80^{o}C. RNAs were extracted from the leaf tissue using the Ambion PureLink Plant RNA Reagent (ThermoFisher Scientific Inc, Waltham, Massachusetts, United States) following the manufacturer's instructions and quantified using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California, United States). Sequence libraries were prepared using either the TruSeq Stranded mRNA Library Prep Kit (Illumina, Sandiego, California, United States) or the KAPA stranded mRNA library preparation kit (Kapa Biosystems, Wilmington, Massachusetts, United States) using the default protocols except for fragmentation at 94°C for 6 min and ten cycles of PCR enrichment. All ten libraries were multiplexed, then *D. binata* and *N. alata* were sequenced together on the same lane of the Illumina HiSeq2000 platform. *Ruprechtia salicifolia* was run on a separate Illumina HiSeq2000 lane with six other samples, *A. robertsoniorum* was run on a separate Illumina HiSeq2500V4 along with ten other samples and *P. auriculata* was run on a separate Illumina HiSeq2500V4 run along with ten other samples (Appendix B).

The raw paired end reads from the newly generated transcriptomes were trimmed and filtered using Trimmomatic (Bolger et al., 2014) with trim settings sliding window 4:5, leading 5, trailing 5 and min length 25. For both *D. binata* and *N. alata*, the three transcriptomes from trap tissues were combined and assembled together. The procedure was conducted as follows: the remaining read set was assembled using Trinity v2.04 (Grabherr et al., 2011) with strand-specific settings and stranded 'RF' and the assembled reads were translated using Transdecoder v2.0 (Haas et al., 2013) guided by BLASTP against a BLAST database consisting of concatenated *Arabidopsis thaliana* and *B. vulgaris* proteome (Dohm et al., 2014), with strand-specific settings. All translated amino acid datasets were reduced with cd-hit v4.6 (-c 0.995 -n 5) (Fu et al., 2012).

Analysis of Sources of Contamination

We tested for within-lane contamination by creating one-to-one ortholog gene trees (using the pipeline described below) and comparing the resulting tree topologies to the expected species tree topology for all samples on the lane. Additionally, we examined *matK* sequences from the assembled transcriptome coding DNA sequence (CDS) data. Using these sequences together with those obtained from GenBank (Appendix B) to represent each of the non-core families used in the analysis, we constructed a phylogeny using maximum likelihood and the settings "-f a -# 200 -m GTRCAT -p 12345 -x 112233" as implemented in RAxML (Stamatakis, 2014). We were unable to recover *matK* from two of the assembled transcriptomes (*A. vesiculosa* and *P. auriculata*), and instead we recovered the *rbcL* gene and ensured that the highest GenBank BLAST hit was that of the same species *A. vesiculosa* (AY096106.1) and *P. auriculata* (EU002283.1) respectively.

Homology Inference and Species Tree Estimation

Homology and orthology inference along with species tree estimation were carried out following Yang and Smith (2014), which is briefly summarized below. The exact commands and programs are available either at at https://bitbucket.org/yangya/phylogenomic_dataset_construction for scripts used in assembling the species tree or https://github.com/jfwalker/JFW_NonCore_Caryophyllales for scripts involved in the downstream analysis. After the peptide and coding DNA sequences were reduced using cd-hit, we created six datasets to explore the influence of taxon sampling and sequence type. Three of the datasets were made using the peptide data. One dataset consisted of all taxa;

one dataset excluded *Ancistrocladus robertsoniorum* and one dataset excluded *Drosophyllum lusitanicum*. We then created corresponding nucleotide sequence datasets with the same taxon content. All steps for the homology inference and species tree estimation were the same for all datasets, except where noted below. The first step was an all-by-all BLASTP search, in the case of the peptide datasets, or an all-by-all BLASTN search in the case of the nucleotide data, which was conducted with an e-value of 10. Putative homolog groups were formed by retaining species with a hit fraction >0.4 and using Markov clustering as implemented in MCL14-137 (Van Dongen, 2000) with the inflation value set to 1.4 and e-value cutoff of 10^{-5}. Only clusters that had at least 4 taxa were retained.

Each cluster was then aligned using MAFFT v7 (Katoh and Standley, 2013) with "--genafpair maxiterate 1000" and trimming of the alignments was conducted using Phyutility v2.2.6 (Smith and Dunn, 2008) with "-clean 0.1". After the alignment we manually checked a random sample of ~10 sequences to ensure high quality alignment. For sequence clusters containing less than 2,000 sequences, the phylogenetic trees were estimated through maximum likelihood as implemented in RAxML v8.2.3 (Stamatakis, 2014) with the model PROTCATWAG (AA) or GTRCAT (DNA). In the case of sequence clusters larger than 2000 sequences, this was done with FastTree 2 (2.1.8) (Price et al., 2010) with the WAG model (AA) or the GTR model (DNA). All single branches greater than 2 substitutions per site were removed as these are likely the result of sequences being pulled together by error or conserved domains. We also removed all terminal branches 10 times or greater in length than their sister branches in the homolog tree for similar reasons. In the case of clades, the analysis took the step-wise average from root to tip and removed it if that was greater than 10 times the length of the sister. Further data refinement was done by removing all the monophyletic tips except the tip associated with the sequence with the highest number of aligned characters after trimming (i.e. most informative) data. The sequence data were then removed from the homolog trees and the process was repeated a second time, to further clean the data.

The support for the homolog trees was analyzed after the second round using the Shimodaira-Hasegawa-like approximate likelihood ratio branch test (Anisimova et al., 2011) as implemented in RAxML, for downstream analysis only branches with (SH-Like => 80) we considered informative. Then one-to-one orthologs were identified from the homolog trees (Yang and Smith, 2014), using *B. vulgaris* and *S. oleracea* as outgroups*,* both of which are in the core Caryophyllales and have genome information. The ortholog trees produced from these methods were then used to extract the amino acid sequence data associated with the given ortholog tree. A dataset was created from one-to-one orthologs containing no missing taxa. Each ortholog produced from each method was then individually aligned using PRANK v.140603 with default parameters (Löytynoja and Goldman, 2008). The alignments were then trimmed using Phyutility with a minimum occupancy of 0.3 being required at each site. Supermatrices were created for all approaches by concatenating all trimmed alignments that had at least 150 characters. A maximum likelihood tree for each supermatrix was estimated using RAxML with the PROTCATWAG model, partitioning by each ortholog group. Node support was evaluated using 200 nonparametric bootstrap replicates. Following this the Maximum Quartet Support Species Tree (MQSST) was found using ASTRAL (v4.10.0) (Mirarab et al., 2014) with default parameters and using the one-to-one ortholog trees as the inputs.

Dating Analysis

To conduct the analysis, we used the "SortaDate" procedure for filtering genes (Smith et al., 2017). In short we took the 1237 orthologs identified in the nucleotide dataset and first found the genes whose gene tree matched the species tree. From the 135 genes that met this criterion, we calculated the variance from each tip to root, using pxlstr from the Phyx package (Brown et al., 2017).The dating analysis was conducted using BEAST (ver. 1.8.3) (Drummond and Rambaut, 2007) on the three genes with the lowest variance as they represent the genes evolving in the most clocklike manner. We used the GTR+G model of evolution and a birth-death tree prior. We calibrated the clade containing the genera *Aldrovanda* and *Dionaea* with a lognormal prior with offset 34 and a mean of 0 and standard deviation of 1 based on a fossil *Aldrovanda* (Degreef, 1997). Because of the low root to tip variance for the three genes (~0.0004), we used the strict clock model for the rates of evolution. We ran the MCMC for 10,000,000 generations and the first 1,000,000 generations were discarded as the burn-in. We summarized the topology as the maximum clade credibility tree. We repeated these analyses using an uncorrelated lognormal clock with an exponential prior and with a lognormal prior to see the influence choice of prior and model had on our analysis.

Gene Family Size Analysis

Two sets of gene families were analyzed, one for the overall largest gene family and one for the gene families previously associated with the adaptation to carnivory in a differential gene expression study (Bemm et al., 2016). To identify the overall largest family, we found the inferred homolog trees that had the largest number of tips, and annotation was done by taking a representative sample from the homolog tree and finding the highest hit on NCBI blast database. For the carnivorous gene families, representative samples from the genes identified in *Bemm et.*

al were downloaded from Genbank (Appendix B). A blast database was created from the downloaded samples and BLASTP was used to identify their corresponding sequences, which were then found in the homologous gene clusters. The number of tips were counted for each homologous gene tree to identify the size of the gene family and number of genes associated with carnivory.

Analysis of Gene Duplications

Gene duplications were analyzed with phyparts (vrs. 0.0.1) (Smith et al., 2015) using the homolog clusters. Only gene duplications with nodes that contained ≥80 SH-Like support were used to identify duplications. The homolog clusters for each of the six datasets were mapped onto their respective species tree topologies. Further analysis of the gene duplications was conducted by finding all gene duplications, irrespective of species tree topology, using a modified version of phyparts. Again in this case only gene duplications that contained (≥80) SH-Like support were removed from the homolog trees. These duplications were then used to create a phylogenetic tree by creating a shared presence matrix from existing duplications and correcting for distance by taking (1/number of shared duplications). The distance matrix was used to create a phylogenetic tree following the Neighbor-Joining method (Saitou N, 1987). The modified version of phyparts and script (GeneJoin.pl) that creates a phylogenetic tree from that output can be found at (https://github.com/jfwalker/JFW_NonCore_Caryophyllales).

Analysis of Gene Tree Conflict

The one-to-one orthologs recovered from the homolog trees were used to analyze the gene tree/species tree conflict at all nodes and this analysis was performed on all six datasets, with their respective gene trees and species tree being used for each individual analysis. The orthologs were all rooted based on outgroups *S. oleracea* and *B. vulgaris* using the phyx program pxrr (Brown et al., 2017). The rooted one-to-one ortholog trees were then compared to the species tree using phyparts with only informative branches being counted. The output of phyparts was used to identify the amount of conflict at each node along with the dominant alternative topology.

Inferring genome duplication events

To infer potential genome duplication events, we visualized the number of synonymous substitutions that were found between the paralogs with all of the taxa. The process was carried out using the script ks_plots.py from Yang et. al 2015 (https://bitbucket.org/yangya/caryophyllales_mbe_2015) which relies upon the pipeline from (https://github.com/tanghaibao/bio-pipeline/tree/master/synonymous_calculation). The pipeline first reduces sets of highly similar sequences using CD-HIT (-c 0.99 -n 5). Following this, an all-by-all BLASTP is carried out within each taxon using an e-value of 10 and -max_target_seq set to 20. The resulting hits with < 20% identity or niden < 50 amino acids are removed. The sequences that have ten or more hits are removed to avoid over representation of gene families. The remaining paralog pairs are then used to infer the genome duplications, as areas where the Ks value is greater than the background rate (Schlueter et al., 2004). First pairwise protein alignments are created using the default setting of ClustalW (Larkin et al., 2007), these are then back translated to codon alignments using PAL2NAL, and the synonymous substitutions rates are calculated using yn00 of the PAML package (Yang, 2007), with Nei-Gojobori correction for multiple substitutions (Nei and Gojobori, 1986).

To infer the phylogenetic locations of genome duplications, we used a comparison of the genome duplication events identified from paralogs mapped onto the Ks plots of multiple species made from the reciprocal blast hits of their orthologs. The process was carried out using the script MultiKs.pl, which can be found at (https://github.com/jfwalker/JFW_NonCore_Caryophyllales). The pipeline works as follows. First the highly similar sequences are reduced using CD-HIT (-c 0.99 –n 5). Then a reciprocal BLASTP is carried out on the peptide transcriptomes where one of the transcriptomes is used as a query and another is used as the database. Following that the top blast hit from the peptide sequences are aligned using MAFFT. The peptide alignment is then matched with the corresponding nucleotide files and the nucleotides are aligned based on the peptide alignment using the phyx program pxaatocdn (Brown et al., 2017). From there the synonymous substitution rates are calculated using yn00 of the PAML package, with the Nei-Gojobori correction for multiple substitutions. The Ks peaks of the genome duplications inferred from the paralogs are then compared to the Ks peaks of the multispecies comparison, if the peak from the single species comparison is smaller than the multi-species, this provides evidence that the genome duplication occurred after the speciation event (Cannon et al., 2015).

Comparing molecular rates among differing gene tree topologies

The gene trees that contained the topologies supporting either *Drosophyllum* and *Ancistrocladus* as sister to all other lineages or *Drosophyllum* and *Ancistrocladus* as sister to *Nepenthes* were identified from the bipartitions removed using the phyx program pxbp (Brown et al., 2017) and the program GeneHybridSplitter.pl (https://github.com/jfwalker/JFW_NonCore_Caryophyllales). The ortholog tree was considered to support *Drosophyllum* and *Ancistrocladus* as the lineage sister to the others if it contained a bipartition containing only *Drosophyllum* and *Ancistrocladus*, a bipartition containing only the carnivorous lineages except *Drosophyllum* and *Ancistrocladus*, and a bipartition containing only and all the carnivorous taxa. The ortholog trees that supported *Drosophyllum* and *Ancistrocladus* sister to *Nepenthes* were identified if the tree contained a bipartition with only *Ancistrocladus* and *Drosophyllum*, a bipartition with both *Nepenthes* species and *Drosophyllum* and *Ancistrocladus*, and a bipartition containing only and all the carnivorous taxa.

The synonymous substitution rates found in both scenarios were calculated using a pairwise comparison of *Drosophyllum* and *Nepenthes alata*, along with a pairwise comparison of *Ancistrocladus* and *N. alata*. The corresponding nucleotide and amino acid sequences of *Drosophyllum* and *N. alata* were removed for all the gene trees that support *Ancistrocladus* and *Drosophyllum* as the basal lineage. The pairwise amino acid sequences were then aligned using MAFFT, and the amino acid alignment was then used to guide the codon based alignment using pxaatocdn. The Ks values for each codon alignment were calculated using the script Ks_test.pl (https://github.com/jfwalker/JFW_NonCore_Caryophyllales), which uses yn00 from the PAML package to obtain the Nei-Gojobori correction for multiple substitutions Ks values. The same procedure for finding synonymous substitutions was then performed on pairwise comparisons of *Drosophyllum* and *N. alata*, where they appear as sister, and was performed on *Ancistrocladus* and *N. alata* for the same situations.

Results

Species tree, dating analyis and gene tree conflict

The monophyly of the non-core Caryophyllales was supported in both the concatenated maximum likelihood supermatrix (Appendix B) and the maximum quartet support species tree

(MQSST) reconciliations (Appendix B), regardless of taxon sampling or molecule type used in the analysis. The divergence of this group based on the use of a strict clock and lognormal prior appears to have occurred ~90 mya ago, with adaptation of carnivory arising ~75 mya (Fig. 3-1): (ucln+logn = 55.9-76.4 mya and ucln+exp = 58.1-110.1 mya). A general trend was that branches of high conflict resulted in shorter branch lengths for both the concatenated supermatrix and the MQSST analysis (Appendix B). A clade of Frankeniaceae and Tamaricaceae was supported as sister to the remaining non-core Caryophyllales in all datasets by most gene trees. In the case of the dataset containing all taxa (AA ALLTAX), the branch supporting this as the lineage sister to everything else showed a large amount of conflict with ~15.4% of genes supporting the topology, ~14.6% supporting a dominate alternate topology of a monophyletic non-carnivorous non-core (NCNC), ~25% supporting other alternate topologies and ~45% of gene trees being poorly supported (SH-Like < 80), with similar results for the five other datasets used to reconstruct the species tree topology. Further support of a non-monophyletic relationship of the NCNC was obtained by looking at the number of uniquely shared gene duplications found by the AA ALLTAX for the families in the carnivorous non-core with the clade of Plumbaginaceae and Polygonaceae was 93. This is in contrast to the three unique gene duplications shared among all members of the NCNC. The MQSST and concatenated ML supermatrix analyses inferred that the next lineage to diverge was a clade containing both the families Plumbaginaceae and Polygonaceae, whose sister relationship received 100% bootstrap support and ~70% genes concordant with the topology with 10.5% conflicting in the case of the AA ALLTAX. This relationship showed up in all datasets regardless of composition of taxa used for the analysis.

All datasets revealed a strongly supported (BS = 100%) clade consisting of the carnivorous families and the non-carnivorous family Ancistrocladaceae. In the case of the AA

ALLTAX dataset the majority of the well-supported gene trees (~57%) were concordant with the species tree topology, with similar results for all other datasets. In all cases, Droseraceae and Nepenthaceae were each monophyletic (Fig. 3-2).

The main discordance in the species tree topology involved the placement of Drosophyllaceae (Fig. 3-2). When all taxa were included Drosophyllaceae was sister to Ancistrocladaceae, a relationship that is well supported by concordant gene signal in both the AA dataset (72.5%) and the CDS dataset (93.7%). However, the placement of the clade containing Drosophyllaceae and Ancistrocladaceae changed depending on sequence type: for AA data it is reconstructed as sister to the Nepenthaceae, whereas for CDS data it is sister to the rest of the carnivorous clade, albeit with no bootstrap support (Fig. 3-2).

When *Ancistrocladus* was excluded from analyses, for both the AA and CDS datasets, Drosophyllaceae appeared as sister to the rest of the taxa in the carnivorous clade (Fig. 3-2b,e). The clade containing Droseraceae and Nepenthaceae has a large amount of discordance with ~18% concordant and 32% conflicting for the AA dataset and ~20% concordant and ~22% conflicting for the CDS dataset. In both cases this was a node where many of the gene trees contained low Shimodaira-Hasegawa-Like support (< 80%). When *Drosophyllum* was excluded from analyses, for both the CDS and the AA datasets, Ancistrocladaceae appeared as sister to Nepenthaceae. Again, the node that defined this relationship had a significant amount of conflict, where in the AA dataset ~25% of the gene trees showed a concordant topology and ~24% showed a conflicting topology. With the CDS dataset ~22% of gene trees were concordant with the species topology and ~24% gene trees were conflicting. Again in both cases many of the gene trees did not have strong SH-Like (≥80) support for either topology.

Analysis of potential hybridization and comparison of synonymous substitutions rates (Ks) between woody and herbaceous species

No differences were found between the synonymous substitution rate between the gene trees supporting the sister position of *Drosophillum lusitanicum* and *Aldrovanda robertsoniorum* to the remaining lineages as opposed to those supporting the two species as sister to only Nepenthaceae (Appendix B). For *D. lusitanicum,* the mean Ks for the trees supporting the sister to the other lineages position was 0.8546, whereas those supporting the position sister to Nepenthaceae had a mean Ks value of 0.8586. In the case of *A. robertsoniorum* those supporting a sister to the other lineages relationship had a mean Ks value of 0.6359 and those supporting a relationship sister to only Nepenthaceae is 0.6358.

Genome duplications and gene family sizes

The single-species Ks plots showed that all the Caryophyllales have at least one peak around 2.0 (Appendix B). These plots also showed one additional peak for all taxa in non-core Caryophyllales except for *A. vesiculosa,* which had two additional peaks, and both *D. lusitanicum* and *Frankenia laevis* did not show any extra peaks. A comparison of Ks values between orthologs and paralogs for species pairs showed that in the case of Plumbaginaceae and Polygonaceae, the genome duplication likely occurred post speciation (Fig. 3-3). This post speciation genome duplication received further support as the two species only shared five unique gene duplications. This same comparison for representative species pairs of Ancistrocladaceae-Nepenthaceae and Droseraceae-Nepenthaceae showed that these genome duplications likely occurred after the divergence of the respective families in each pair (Fig. 3-3).

An among Droseraceae comparison showed the duplication to have occurred after speciation in *Dionaea* but before speciation in *Drosera* (Appendix B). The peak for the duplication appeared to be before-speciation in a comparison to *Drosera* and *Aldrovanda* (Appendix B). Overall, the shared unique gene duplications and Ks plots support the inference of seven separate genome duplications across the non-core Caryophyllales, with six occurring after divergence of the families and none being uniquely shared by any two families in the group (Fig. 3-3).

An analysis of the size of homologous gene families on the AA ALLTAX dataset showed that the largest gene family consisted of 3498 homologs (Appendix B) and this family was associated with the function "putative leucine-rich repeat receptor-like protein kinase". When further broken down into genes that are associated with carnivory, we found that the largest of these gene families was the "Plant Peroxidase" family (Appendix B). On average, we did not find any specific gene family to have a disproportionate number of duplicated genes in the carnivorous plants as compared to the rest of the samples in the remaining non-core Caryophyllales, however, the plant peroxidase family has shrunk in the carnivorous lineage.

Contamination checking and homology and orthology inference

Three major steps were taken to ensure that we would minimize the possibility of contamination in our samples. The first step was to extract the RNAs, prepare the sequencing libraries, and sequence the samples on separate lanes at different times. This was done for all samples we processed in this study other than *Nepenthes alata*, *Drosera binata*, and the previously published *D. lusitanicum*, which were sequenced together on a single lane. The next step was to create one-to-one ortholog phylogenetic trees out of the samples that were on the same lane, which showed most gene trees support previously accepted hypotheses for the often distantly related species on the lane. The final step was to ensure that the *matK* sequence from each of our assembled transcriptome shared the closest evolutionary relationship with a *matK* sequence taken from the same genus for each sample (Appendix B).

The datasets were made of the following taxon compositions for both amino acid (AA) and coding DNA sequence (CDS): all 13 taxa included (ALLTAX), all taxa except *D. lusitanicum* (NO DROS), and all taxa except *A. robertsoniorum* (NO ANC). The two datasets with all 13 taxa revealed that the inferred number of homolog clusters containing at least four taxa was the greatest using nucleotide data (Appendix B). This is in contrast with both datasets that consisted of 12 taxa, in which the amino acid datasets inferred more homolog clusters than the nucleotide datasets. The complete taxa one-to-one orthology inference was comparable between all datasets of different taxa composition, where each time the amino acid dataset detected roughly 400 more one-to-one orthologs than its corresponding nucleotide dataset (Appendix B).

Discussion

Discordance among species trees and gene trees

Our transcriptome data confirm the monophyly of the carnivorous clade of Caryophyllales detected in previous studies (Meimberg et al., 2000; Brockington et al., 2009) and imply an ancient origin for the group, which our analyses suggest originated between 68-83 mya (Fig. 3-1). Our analyses further confirm that carnivory was the likely ancestral character state for the carnivorous clade, and that a mucilage trap characterized the progenitor of this clade (Heubl et al., 2006). Nevertheless, the subsequent evolution of life history within the carnivorous clade is less certain because it depends upon the topology of the earliest branches within the group, which have been unstable in previous analyses (Meimberg et al., 2000; Brockington et al., 2009; Hernández-Ledesma et al., 2015).

The large datasets generated in our study provide unique insight into the sources of this topological instability (Galtier and Daubin, 2008). For example, the shifting phylogenetic placement of *D. lusitanicum* could result from events such as horizontal gene transfer, incomplete lineage sorting, and/or ancient hybridization between an ancestral lineage that diverged prior to the other carnivorous Caryophyllales and one that diverged after the speciation event between Ancistrocladaceae and Nepenthaceae. The Nepenthaceae provides a logical source of hybridization as many of the species in genus are still capable of producing viable hybrids and do so in the wild (McPherson, 2009). If hybridization were the cause, we would expect two points of coalescence between *D. lusitanicum* and *N. alata* that would be associated with different synonymous substitution (Ks) values, as they would be influenced by the amount of time there was shared common ancestry with *N. alata*. An examination of Ks values did not reveal a difference in Ks values between the gene trees supporting the sister to all other lineages position or the sister to only Nepenthaceae position from the nucleotide data for either *D. lusitanicum* or *A. robertsoniorum* (Appendix B). This provides some evidence that something other than hybridization may be the cause. However, full genome sequences would be necessary to improve confidence in our ability to discriminate among these processes because they would allow for direct association of phylogenetic signal over contiguous regions of chromosomal space (Fontaine et al., 2015). However, we did find that Ks values varied greatly between the *D. lusitanicum* and *A. robertsoniorum* comparisons, which may result from differences in habit, with the lineage of *Ancistrocladus* + Dioncophyllaceae transitioning to lianas and *Drosophyllum* retaining the ancestral herbaceous life history (Smith and Donoghue, 2008; Yang et al., 2015).

The remaining families of non-core Caryophyllales (Polygonaceae, Plumbaginaceae, Tamaricaceae, and Frankeniaceae) have previously been inferred to be a clade (Meimberg et al., 2000; Brockington et al., 2009; Soltis et al., 2011; Hernández-Ledesma et al., 2015; Yang et al., 2015), but our transcriptome-based analyses suggest that the clade of Frankeniaceae and Tamaricaceae and that of Plumbaginaceae and Polygonaceae are successively sister to the carnivorous clade. It is possible that this conflict is the result of our study including more informative phylogenetics characters in the analysis. However, it may also be the result of our relatively limited taxon sampling for these families and/or from the large number of conflicting gene trees associated with divergence events among these three groups (Fig. 3-1). The large number of conflicting gene trees may, itself, be the result of ILS associated with the relatively rapid divergence of these groups, as demonstrated by the short branch lengths from the MQSST analysis and concatenated supermatrix analysis (Appendix B). The 93 uniquely shared gene duplications provide evidence for the sister relationship between the carnivorous clade and the clade of Plumbaginaceae + Polygonaceae as a clade consisting of all the NCNC only contains 3 uniquely shared gene duplications. However, it should be taken into account that the higher number of gene duplications shared between Plumbaginaceae, Polygonaceae and the carnivorous Caryophyllales could be the result of biased sampling that is inherent when not retrieving all coding genes, as transcriptomes are typically only found to recover up to half of coding genes (Yang and Smith, 2013). This provides a potentially biased sample for data when looking at uniquely shared gene duplications.

The disagreement between the supermatrix and MQSST methods of species tree reconciliation was likely a product of how the genes were treated in the analyses. In the MQSST all genes are given equal weight regardless of their influence and strength of the phylogenetic signal provided by the characters that created them, whereas in the supermatrix approach more influential genes provide a stronger signal for the overall matrix. Recent work has shown that the topology of a phylogeny may change greatly by the influence of just a couple genes and is especially influential at nodes of high conflict (Brown and Thomson, 2016; Shen et al., 2017; Walker et al., 2017). The conflicting node for the CDS topology, however, received no bootstrap support.

Our results help to illustrate the important role that taxon sampling plays even when using character-rich datasets such as those used in phylogenomic reconstructions. In the analyses presented here, *D. lusitanicum* changed positions depending on the sampling used (Fig. 3-2). This discrepancy was not identified by the non-parametric bootstrap method, as 100% support was given to all nodes in all the reconstructions using the amino acid datasets, regardless of the position of *D. lusitanicum*. This helps to emphasize the importance of looking at more than just the non-parametric bootstrap in phylogenomic reconstructions, as in our datasets it is prone to Type I error and using transcriptome data allows us to examine conflicting signals. The non-parametric bootstrap, however, provided no support for the conflicting signal produced from nucleotide data. While we are unable to include Dioncophyllaceae in our analyses because of the difficulty in obtaining tissue, it is unlikely that inclusion would dramatically change carnivorous relationships given the strong support for its sister relationship to Ancistrocladaceae in all previous analyses (Heubl et al., 2006; Brockington et al., 2009).

At least seven independent paleopolyploidy events in a group of less than 2500 species

Over the past decade, ever-larger phylogenomic datasets and improved methods for detecting genome duplications have revealed that paleopolyploidy is much more common in plants than previously thought (Barker et al., 2008, 2016; Yang et al., 2015). Previous evidence has suggested that the non-core Caryophyllales contain at least three paleopolyploidy events (Yang et al., 2015). Genome duplications have previously been implicated to be a source of novelty (Freeling and Thomas, 2006; Edger et al., 2015), a source of increased diversification (Tank et al., 2015), and decreased diversification (Mayrose et al., 2011). The seven inferred genome duplications of our analysis indicate that genome duplication has been a common occurrence in the history of the non-core Caryophyllales and is especially prevalent considering the group is estimated to have less than 2500 species (Soltis et al., 2006). Our results also support a shared genome duplication between the core and non-core Caryophyllales giving support to the evidence that at least one duplication is shared by the entire clade (Dohm et al., 2012). From our dataset it appears most of the non-core Caryophyllales families have unique genome duplication events. We found a discrepancy in the location of the duplication when comparing *Drosera* to *Dionaea* and when comparing *Drosera* to *Aldrovanda*. This may be due to the duplication occurring shortly before speciation or to the difference in rates of evolution found between *Aldrovanda* and *Dionaea* (Appendix B). Without exhaustive sampling of each family it will not be possible to pinpoint the phylogenetic locations of the putative duplication events and hence it is not currently possible to determine whether a given paleopolyploid event acted to drive speciation and/or promote ecophysiological and morphological novelty. Nevertheless, the rich diversity and large number of genome duplications present within the non-core Caryophyllales suggests that this group will be a powerful tool for understanding genome and phenome evolution.

Acknowledgements

We thank Edwige Mayroud, Joseph Brown, and Oscar Vargas for thoughtful comments on the manuscript and Ning Wang, Sonia Ahluwalia, Jordan Shore, Lijun Zhao, Alex Taylor and Drew Larson for helpful discussion on the manuscript; M. Raquel Marchán Rivadeneira for help with lab work; and Deborah Lalumondier and Justin Lee at the Missouri Botanical Garden for access to their living collections. We would also like to thank the Associate Editor and two anonymous reviewers for comments that helped improve the manuscript. The molecular work of this study was conducted in the Genomic Diversity Laboratory of the Department of Ecology and Evolutionary Biology, University of Michigan. This work was supported by NSF DEB awards 1352907 and 1354048.

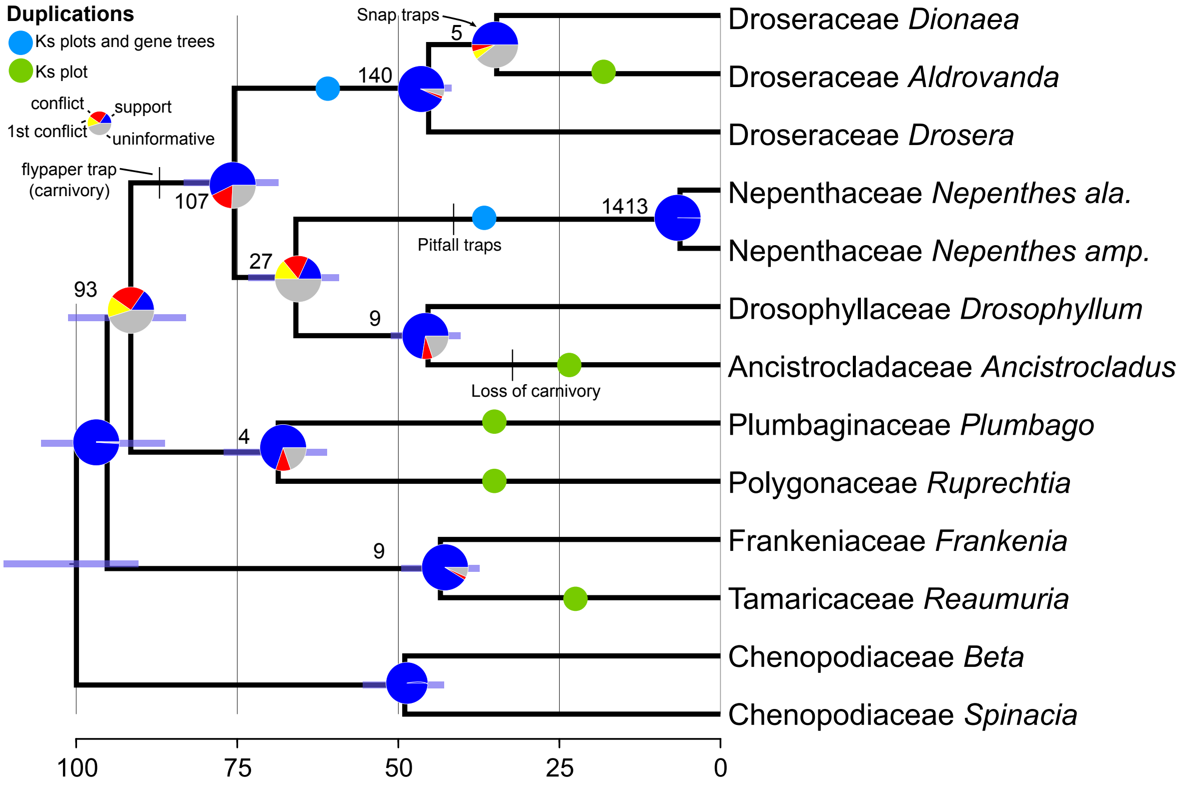


Figure 3-1 Inferred and dated species tree from the three-gene Bayesian dating analysis

Numbers on each branch represent inferred shared unique to clade gene duplications, and branch lengths are proportional to time. Circles on branches represent inferred genome duplications, position supported only by Ks plots (Green) and position supported by Ks plots along with shared gene duplications (Blue). Pie charts show gene tree conflict evaluations at each node, proportion concordant (Blue), proportion conflicting (Red), dominant alternative topology (Yellow) and unsupported with SH-Like less than 80 (Grey). Ancestral states on branches taken from *Heubl et. al 2006*.

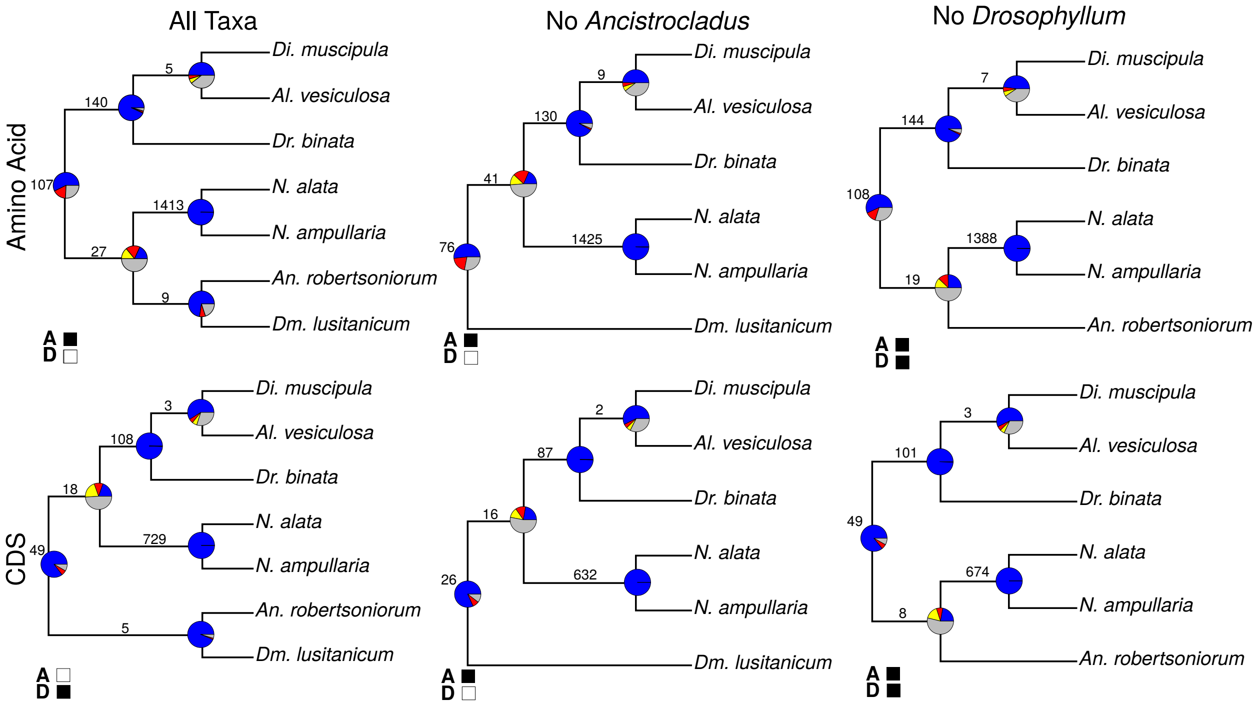


Figure 3-2 The influence of taxon sampling and sequence type on inferred tree topology

Respective topologies are from the RAxML supermatrix analysis, filled boxes are used to represent concordance with a different method of species tree reconciliation "A" represents Astral (MQSST) and "D" represents Distance matrix reconstruction. Star near the node indicates BS support of 0, all other nodes have BS support of 100. Numbers on each branch represent inferred gene duplications. Pie charts show gene tree conflict evaluations at each node, proportion concordant (Blue), proportion conflicting (Red), dominant alternative topology (Yellow) and unsupported with SH-Like less than 80 (Grey).

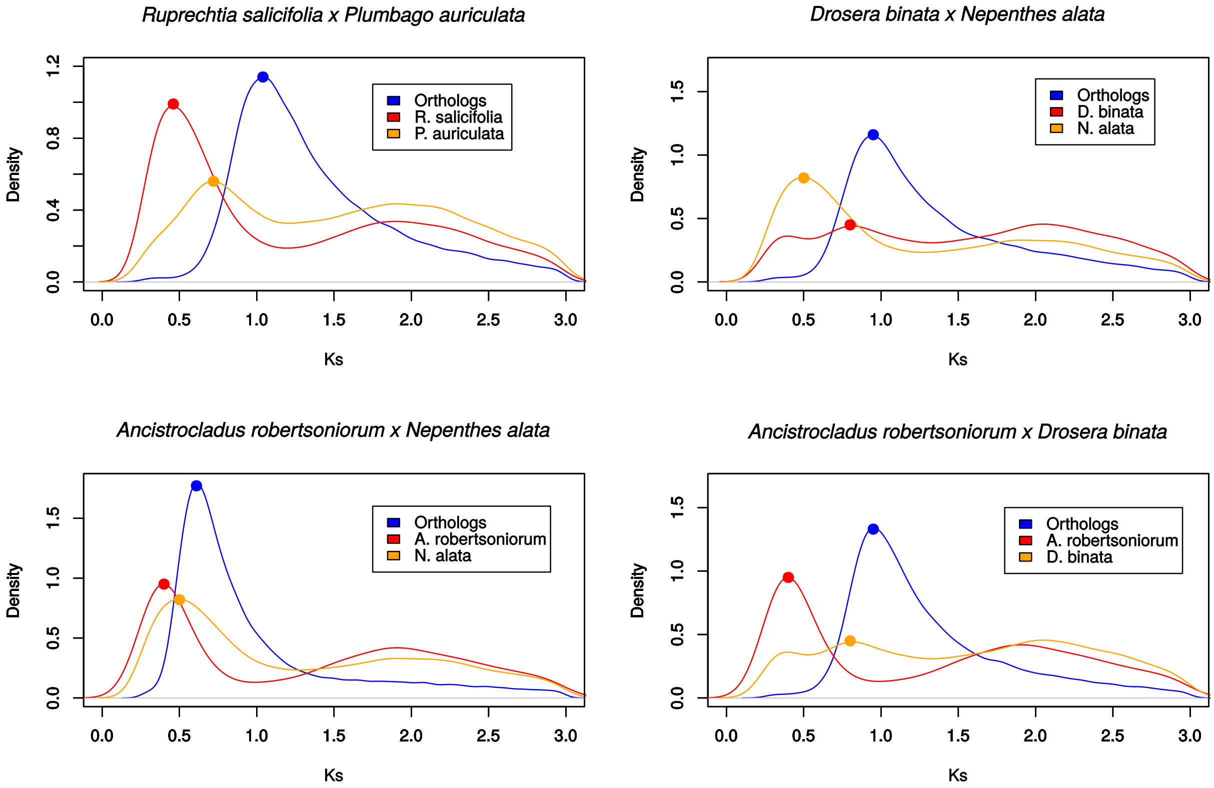


Figure 3-3 Representative Ks plots

Density plots representing the peak of the Ks values inferred from reciprocal orthologs (Blue) and those inferred from the within species paralogs (Red and Orange), with the density calculated for Ks values (=>0.25).

Chapter IV

From cacti to carnivores: Improved phylotranscriptomic sampling and hierarchical homology inference provide further insight into the evolution of Caryophyllales

Preamble: This chapter is our manuscript that has been published, the citation for this manuscript is: JF Walker, Y Yang, T Feng, A Timoneda, J Mikenas, V Hutchison, C Edwards, N Wang, S Ahluwalia, J Olivieri, N Walker-Hale, LC Majure, R Puente, G Kadereit, M Lauterbach, U Eggli, H Flores-Olvera, H Ochoterena, SF Brockington, MJ Moore and SA Smith. From cacti to carnivores: Improved phylotranscriptomic sampling and hierarchical homology inference provides further insight to the evolution of Caryophyllales (2018). ***American Journal of Botany***, 105:3.

Abstract

The Caryophyllales contains ~12,500 species and is known for its cosmopolitan distribution, convergence of trait evolution, and extreme adaptations. Some relationships within the Caryophyllales, like those of many large plant clades, remain unclear and phylogenetic studies often recover alternative hypotheses. We explore the utility of broad and dense transcriptome sampling across the order for resolving evolutionary relationships in Caryophyllales. We generated 84 transcriptomes and combined these with 224 publicly available transcriptomes to perform a phylogenomic analysis of Caryophyllales. To overcome the computational challenge of ortholog detection in such a large data set, we developed an approach for clustering gene families that allowed us to analyze >300 transcriptomes and genomes. We then inferred the species relationships using multiple methods and performed gene tree conflict analyses. Our phylogenetic analyses resolved many clades with strong support, but also showed significant gene-tree discordance. This discordance is a common feature of phylogenomic studies but also represents an opportunity to understand processes that have structured phylogenies. We also found taxon sampling influences species-tree inference, highlighting the importance of more focused studies with additional taxon sampling. Transcriptomes are useful both for species tree inference and for uncovering evolutionary complexity within lineages. Through analyses of gene-tree conflict and multiple methods of species tree inference, we demonstrate that phylogenomic data can provide unparalleled insight into the evolutionary history of Caryophyllales. We also discuss a method for overcoming computational challenges associated with homolog clustering in large datasets.

Introduction

The Caryophyllales [*sensu* Angiosperm Phylogeny Group IV (APG, 2016)] contain an estimated ~12,500 species and are found on all continents and in all major terrestrial ecosystems (Hernández-Ledesma et al., 2015). The clade is notable not only for its diversity and broad ecological and geographic distribution but also for its array of unique morphological and ecophysiological adaptations. Many Caryophyllales (most famously many cacti) are noted for their extreme drought tolerance, but the clade also contains species that exhibit extreme cold tolerance (Cavieres et al., 2016), halophytism (Flowers and Colmer, 2008; White et al., 2017), heavy metal hyper-accumulation (Moray et al., 2016), carnivory (e.g. Venus flytrap, sundews, and *Nepenthes* pitcher plants) (Albert et al., 1992; Givnish, 2015), betalain pigmentation

(Brockington et al., 2015), C_4 and CAM photosynthesis (Wang et al., In review; Sage et al., 2011; Moore et al., 2017; Sage, 2017), and succulence (Sajeva and Mauseth, 1991; Eggli and Nyffeler, 2009). Most of these adaptations are known to have arisen multiple times throughout the clade, making Caryophyllales a key natural laboratory for understanding trait evolution in angiosperms. The clade also includes numerous economically important species (e.g., beets, quinoa, and spinach), bolstering its utility as a model system for understanding morphological and physiological evolution.

Previous phylogenetic work, focused on resolving the backbone relationships of Caryophyllales, has utilized morphology (Rodman et al., 1984), targeted gene sequencing (Rettig et al., 1992; Brockington et al., 2009, 2011; Schäferhoff et al., 2009), plastome sequencing (Arakaki et al., 2011), and transcriptome data (Yang et al., 2015; Yang et al., 2017). These studies have resulted in the expansion of the traditional Caryophyllales (i.e., corresponding essentially with the original Centrospermae) to include other families (e.g., Polygonaceae, Plumbaginaceae, Droseraceae, Rhabdodendraceae) and the recircumscription of a number of families, especially the division of previously broadly circumscribed Molluginaceae, Phytolaccaceae, and Portulacaceae APG, 2016. These taxonomic rearrangements have resulted in the 38 families currently recognized by APG IV (2016) as well as the more recently proposed Corbichoniaceae (Thulin et al., 2016). Almost all of these families have been shown to be monophyletic, with the possible exception of Phytolaccaceae due to the uncertain position of the tropical liana *Agdestis clematidea* (Hernández-Ledesma et al., 2015). Our understanding of relationships among these families has advanced greatly during the past 20 years. For example, there has been consistent support at the base of the extant Caryophyllales for a split between the non-core Caryophyllales, consisting of the carnivorous families (Droseraceae, Drosophyllaceae, Nepenthaceae, Ancistrocladaceae, Dioncophyllaceae) and allies (Tamaricaceae, Frankeniaceae, Polygonaceae, and Plumbaginaceae), and a larger clade containing the remaining diversity of the order (Brockington et al., 2009; Hernández-Ledesma et al., 2015). Within the latter clade, there is support for a grade composed of four species-poor families (Rhabdodendraceae, Simmondsiaceae, Asteropeiaceae and Physenaceae) that leads to a well-supported clade containing all of the core members of Caryophyllales (i.e., the old Centrospermae) (Hernández-Ledesma et al., 2015). The diversification within several clades was apparently very rapid (Arakaki et al., 2011), making resolution of the backbone phylogeny of this clade difficult (Hernández-Ledesma et al., 2015). The use of genome data (Jarvis et al., 2014; Fontaine et al., 2015), RADSeq (Eaton et al., 2016), genotyping-by-sequencing (Fernández-Mazuecos et al., 2017), and transcriptome data (Dunn et al., 2008; Smith et al., 2011; Cannon et al., 2015; Pease et al., 2016) have all proven to be robust tools for inferring recalcitrant evolutionary relationships at both shallow and deep time scales, but to date these tools have not been applied to Caryophyllales with sufficient taxon sampling to test hypotheses of early-diverging relationships.

Transcriptomes hold considerable promise as a phylogenetic tool as they provide a relatively cost-effective way to generate a wealth of sequence data for evolutionary analyses, including the exploration of gene-tree conflict and gene/genome duplications (Wickett et al., 2014; Cannon et al., 2015; Smith et al., 2015; Yang et al., 2017). For example, in a study using 92 transcriptomes to reconstruct land-plant relationships, Wickett et al. (2014) demonstrated that phylotranscriptomic data sets provide highly informative data for resolving deeper-level phylogenetic relationships but some relationships were sensitive to reconstruction method. Underlying these sensitive relationships is often gene tree conflict that may arise from a variety of biological causes, including but not limited to incomplete lineage sorting (ILS), hybridization, hidden paralogy, and horizontal gene transfer (Galtier and Daubin, 2008). Gene tree conflict makes it difficult to assess species relationships, as phylogenetic hypotheses are the product of the genes selected for an analysis (Maddison, 1997; Rokas et al., 2003; Walker et al., 2014; Smith et al., 2015), and individual genes can have overwhelming influences on the species-tree topology in phylogenomic data sets (Brown and Thomson, 2017; Shen et al., 2017; Walker et al., 2017). Large multi-locus data matrices may also result in artificially inflated support (Seo, 2008), masking underlying conflict. Futhermore, taxon sampling can affect phylogenetic reconstruction using both coalescent and supermatrix methods (Wickett et al., 2014; Walker et al., 2017). These problems, however, are not a consequence of using transcriptomes per se—rather, transcriptome analyses have exposed problems that have always been present but have been overlooked due to limited data sets. In short, the use of transcriptome data sets provides novel insights into evolutionary history and leads to biological insights that are not obtainable from a handful of loci (Yang et al., 2015; Pease et al., 2016; Smith et al., 2017).

We explore the conflict underlying relationships across the phylogenetic backbone of Caryophyllales using a dataset consisting of 295 transcriptomes and 3 genomes, collectively comprising 32 of the 39 families of Caryophyllales (Byng et al., 2016; Thulin et al., 2016). Due to the severe computational burden imposed by the exponential scaling of all-by-all BLAST during homolog detection, we outline a method of homolog clustering through post-order traversal (tip-to-root). This allowed us to conduct the all-by-all procedure on individual clades that are then combined in a hierarchical manner. Our analyses highlight the tremendous power of using large datasets for inferring species relationships, but they also reveal some of the limitations of large phylogenomic analyses for species relationship inference.

Materials and Methods

Data availability

The raw reads for transcriptomes generated for this study have been deposited in the NCBI sequence read archive (Bioproject SRP127816). Assemblies, orthologous gene clusters, alignments, and trees are available on Dryad (https://doi.org/10.5061/dryad.470pd). Scripts and programs written for this project can be found at Bitbucket (https://bitbucket.org/jfwalker/ajb_bigtree).

Taxon Sampling, Tissue Collection, Sequencing, and Read Assembly

Taxon sampling was designed to broadly cover Caryophyllales. In total, our sampling includes 295 Caryophyllales transcriptomes and three Caryophyllales genomes, representing 298 species and 32 of the 39 families in the clade; the phylogenetic distribution of the species sampled is shown in a collapsed genus level tree of (Smith et al., 2017) (Fig. 4-1). The families Asteropeiaceae, Barbeuiaceae, Corbichoniaceae, Dioncophyllaceae, Halophytaceae, Lophiocarpaceae, and Rhabdodendraceae were not sampled due to the difficulty of obtaining fresh tissue of these taxa. We also included *Agdestis clematidea* to test its phylogenetic position within the phytolaccoid clade (Nyctaginaceae, Petiveriaceae, Phytolaccaceae s.l., Sarcobataceae). We sampled ten outgroups spanning the asterids (*Mimulus guttatus, Solanum lycopersicum, Ilex paraguariensis, Actinidia deliciosa, Vaccinium corymbosum, Camptotheca acuminata* and *Davidia involocrata*), rosids (*Vitis vinifera*), Ranunculales (*Aquilegia coerulea*), and Santalales (*Taxillus nigrans*). A summary of all 84 newly generated transcriptomes can be found in the dryad, along with the sources of the data for previously generated transcriptom. In many cases a previously assembled transcriptome was used, in which case the Dryad repository where that assembly was downloaded is listed.

The 84 newly generated transcriptomes were sequenced and processed following the previously developed phylotranscriptomic workflow (Yang et al., 2017). In short, RNA was obtained from fresh tissue that was flash frozen in liquid nitrogen and stored at -80°C. When possible, the RNA extraction was carried out using a mixture of both young leaf and flower bud. The various methods used for the newly generated transcriptomes can be found in the dryad. All RNA-seq libraries were stranded to simplify assembly and translation. Paired-end sequencing for all newly generated transcriptomes was performed using Illumina HiSeq platforms. Sequence assembly and translation were conducted using previously designed protocols as outlined in Brockington et al., 2015; any differences are highlighted in Walker et al. 2018.

Construction of species trees

We conducted two analyses to reconstruct the relationships within the Caryophyllales. In the first, we conducted a hierarchical clustering method across the entire Caryophyllales (abbreviated ALL throughout), and in the second, we conducted targeted analyses on each well-sampled major group (abbreviated IND throughout).

Reconstruction of the Caryophyllales species tree with hierarchical clustering (ALL)

Tip clustering—The code developed and used for this project can be found at (https://bitbucket.org/jfwalker/ajb_bigtree) and the overarching procedure of tip-to-root clustering has been incorporated into PyPHLAWD (Smith and Brown, 2018). This method utilizes a taxonomy tree based on previous phylogenetic hypotheses. Homologs were first clustered by binning transcriptomes within taxonomic families (which we refer to as tip clustering), and clustering then worked backward toward the root of a taxonomy tree (internal node clustering; 4-2.). Hence, taxonomic families were the tips for the post-order clustering (Fig. 4-2). The family Amaranthaceae was separated into Chenopodiaceae and Amaranthaceae (Hernández-Ledesma et al., 2015). *Agdestis clematidea* was treated as its own tip within the monotypic Agdestidaceae (see Results) and was clustered with the monotypic families due to its conflicting phylogenetic positions (Hernández-Ledesma et al., 2015). The addition of these two families into the analysis expands the total Caryophyllales sampling to 34 families and the total possible families for Caryophyllales to 41 (i.e., the 38 families recognized by APG IV, plus Corbichoniaceae, Agdestidaceae, and Chenopodiaceae). The analysis was conducted on 19 bins of families with 3 or more species represented, 1 bin for all families with less than 3 species represented, and one bin for the outgroups, for a total of 21 bins (Fig. 4-2). The size of these bins ranged from as many as 39 individual species in Caryophyllaceae and Chenopodiaceae, to as few as three in a variety of families (Fig. 4-2). The first step for all transcriptomes and genomes was to reduce sequence redundancy in the translated amino acid data sets using cd-hit (-c 0.995 –n 5) (Fu et al., 2012). Clades including three taxa or more were clustered into putative homolog groups following (Yang and Smith, 2014); The method consists of conducting an all-by-all BLASTP (Altschul et al., 1997), with an E-value cutoff of 10. The top 1000 hits were retrieved and putative homolog groups were retained for species clusters with a hit fraction >0.4. Subsequently, Markov clustering was conducted as implemented in mcl (Van Dongen, 2000), with the inflation value cutoff set to 1.4 and the E-value cutoff set to 10^{-5} "–abc –te 18 –tf 'gq(5)' ".

Resulting phylogenetically informative clusters (≥4 sequences) were separated out for further filtering and remaining clusters (<4 sequences) being retained for node level clustering.

Clusters with four or more sequences were then aligned using MAFFT v7 (Katoh and Standley, 2013), conducted for 1000 cycles of iterative refinement, with the setting "–auto –amino –maxiterate 1000". The alignments were then trimmed for 10% column occupancy using the phyx (v.0.99) program pxclsq (Brown et al., 2017) with the settings "-p 0.1 –a". After each approximation, roughly 10 homolog clusters were manually checked to ensure the alignment and cleaning procedures were performed properly. Phylogenetic trees were then estimated for each potential homolog cluster through maximum likelihood; this was performed using the RaxML v8.2.3 (Stamatakis, 2014) algorithm (for <100 sequences) and the FastTree2 v2.1.8 (Price et al., 2010) algorithm (for >100 sequences). In both cases the trees were estimated under the WAG model of protein evolution.

Each inferred homolog cluster then had all putatively spurious tips filtered out. This was accomplished by removing tips based on relative and absolute branch length criteria outlined by Yang and Smith 2014. The absolute tip cutoff used was 3 substitutions per site and the relative tip cutoff was 2 substitutions per site. These values were used because anything of that length or greater likely represented poor alignment or some form of long-branch attraction based upon conserved domain regions and could lead to compounding issues in downstream alignment and tree inference. The homolog tree was analyzed for all clades that consisted solely of genes from the same taxa; these were then condensed down to a single tip, which was chosen based on the criterion of having the most potentially informative sites (i.e. most amino acids in trimmed alignment). The condensing of these clades was carried out, because any clade consisting solely of tips from the same individual was likely the product of different isoforms or in-paralogs, neither of which provides a means of inferring species relationships. The sequences of the remaining tips were then extracted to form new homolog clusters, with which the same process was again performed two more times for further refinement. The bin containing all small families and the bin containing all outgroups were separately combined and clustered using the same method as the individual families.

Internal node clustering—Clustering at internal nodes of the taxonomy tree was conducted using a post-order tree traversal method (tip-to-root), which was performed following the predicted topology from the Angiosperm Phylogeny Website (Stevens, 2015) (Fig. 4-2) which itself represents a continuously updated compilation of previously inferred phylogenies (e.g. Cuenod et al., 2002; Brockington et al., 2009; Christenhusz et al., 2014). The method proceeds by using the pre-clustered groups generated by the tip clustering step. The predicted sister tips are the first to be combined (e.g., Cactaceae and Portulacaceae; Fig. 4-2). The combination occurs by first creating a BLAST database from one of the tips (or a node depending on where the clustering is occurring). This database consists of random representatives from each of the clustered homologous genes. The number of random representatives was determined by the size of the homologous gene cluster. For clusters with fewer than four sequences, all sequences from the cluster were used; for clusters with four or more sequences, 4 + sqrt (# sequences in the cluster) were randomly selected and added to the database to allow for proportional representation of the cluster.

After the database was initiated from one tip, a BLAST analysis was performed for representative sequences from the other tip, with the representatives being chosen based on the same criteria. The BLASTP analysis was conducted using an e-value cutoff of 1e-3 and only the top hit was retrieved. All clusters from one sister tip/node were then combined with their top hit from the other sister tip/node, using a one-sided BLAST approach. If multiple hits occurred between the two then the new node cluster was formed consisting of all homologous gene families that had a hit.

For example, in the case of Portulacaceae and Cactaceae, the new node level cluster "Cactaceae+Portulacaceae" theoretically could contain all 44 representative taxa from those two families. The next step for the inferred homologs at the node level "Cactaceae+Portulacaceae" bifurcate is to combine Anacampserotaceae, with the newly formed homolog cluster of (Cactaceae+Portulaceae) labeled "1" on Fig. 4-2, which in turn would form the cluster Cactaceae+Portulaceae+Anacampserotaceae, labeled "6" on Fig. 4-2. In later steps, only clusters with less than 5000 sequences are retained as future tree building and alignment steps often have issues with such large data sets. An outline of when this occurs can be found in Fig. 4-2.

Although predominantly conducted in a post-order means, or from tip-to-root, the procedure included some deviations. After that the cluster containing the non-core Caryophyllales, single families, and outgroups was then combined with the core Caryophyllales (internal node 20, Fig. 4-2). This method results in a significant decrease in computational burden imposed by large homolog groups, but due to the removal of clusters smaller than 5000 sequences also causes the final homolog clusters to be smaller than those usually produced by an all-by-all BLAST.

Inference of final gene trees—After the formation of homolog clusters, inference of the final gene trees was conducted by first aligning with MAFFT and trimming the aligned matrix with pxclsq with the settings described above. In the first round of gene-tree inference, FastTree2 v2.1.8 was used with the same settings as noted above to infer all individual gene trees. Next, all sequences with an absolute branch length of two substitutions/site and a relative branch length of one substitution/site were trimmed. Furthermore, any clades that consisted of only genes from a single taxon were again trimmed down to only the gene with the highest number of aligned characters. Next any clade including genes from at least four taxa as well as a branch with at least 1 substitution/site was split into a separate homolog group. The same process was then repeated to help further refine the data set.

Orthologous sequences were inferred from the inferred homologous gene trees using the Rooted Tree (RT) method (Yang and Smith, 2014) and specifying *Aquilegia caerulea* (Ranunculaceae)*, Taxillus nigrans* (Loranthaceae)*,* and *Vitis vinifera* (Vitaceae) as outgroups with a minimum of 50 sequences required as ingroup taxa. The specification of three outgroups in the RT method meant that the final tree contained 305 out of the 308 taxa used in the analysis. The other outgroup taxa were kept as ingroups as they are predicted to form a clade with Caryophyllales and needed to root the species tree after final inference. The orthologous genes were then aligned using MAFFT with the settings above, cleaned with pxclsq for 30% column occupancy, and only alignments that still contained at least 150 characters were retained after cleaning. Gene trees were then estimated using RAxML and tips longer than 0.8 subs/bp were removed and any internal branches longer than 0.8 subs/bp or greater were separated. Then clades with fewer than 50 sequences or fewer than 17 different families were removed from the species tree analysis. The resulting set of 1238 gene trees was used for the downstream MQSST species tree analysis. Finally, we filtered for genes that contained at least 17 different families and 200 taxa, resulting in 58 orthologs for the supermatrix analysis.

Species tree inference for Caryophyllales (ALL)—We inferred a species tree for the dataset including all of the Caryophyllales with two methods. In the first, we conducted a maximum likelihood analysis, as implemented in RAxML v8.2.3, on a supermatrix of 58 orthologs concatenated using pxcat (from phyx; Brown et al., 2017). The supermatrix was partitioned by ortholog, with the WAG substitution model specified for each partition; final inference was conducted using Γ rate variation (PROTCATWAG in RAxML). Support for the tree was evaluated by running 100 rapid bootstraps as implemented in RAxML and for 200 replicates of the quartet sampling method (Pease et al., 2017). The second method employed the Maximum Quartet Support Species Tree (MQSST) algorithm as implemented in ASTRAL-II (v.4.10.12) (Mirarab and Warnow, 2015). This was conducted using the 1238 orthologs that contained at least 17 families and 50 taxa. Support for the tree was inferred using local posterior probabilities (Sayyari and Mirarab, 2016).

Reconstruction of densely sampled clades within Caryophyllales with individual analyses (IND)

Although our hierarchical clustering method was effective in overcoming the computational challenge in orthology inference, taxon sampling may affect orthology inference due to a variety of reasons (e.g., heterogeneity in evolutionary rates, gene/genome duplication, etc.). As such we also conducted species-tree analyses on the five individual clades of interest to help verify the species relationships obtained from using the 305 taxa dataset. The densely sampled clades we chose to analyze separately included Nyctaginaceae, Caryophyllaceae, Amaranthaceae+Chenopodiaceae, Cactaceae, and the clade of non-core Caryophyllales. The methods and settings used for tree inference in each case varied, given the heterogeneity in evolutionary rates across each of the separate clades; therefore, we have outlined settings and modifications below. All statistics, as reported by pxlssq (from phyx; Brown et al. 2017) for the final matrices, can be found in Appendix C.

Caryophyllaceae—Clustering was performed using the same methods as the tip level clustering, but included three Chenopodiaceae (*Spinacia oleracea*, *Chenopodium quinoa*, and *Beta vulgaris*), two Amaranthaceae (*Alternanthera brasiliana* and *Tidestromia lanuginosa*), and *Achatocarpus gracilis* (Achatocarpaceae) as outgroups. The homolog trees then had spurious tips trimmed using an absolute cutoff of 2 substitutions/site and the monophyletic tips were then masked leaving the tip with the most aligned characters. Orthologs were identified using Maximum Inclusion (MI) (Yang and Smith, 2014). Of these identified orthologs, groups containing at least 40 of the 45 taxa were chosen, resulting in 999 inferred orthologs. The individual orthologs were then aligned with MAFFTv7, with the settings "--auto --amino --maxiterate 1000", and alignment trimmed for 10% minimum occupancy using pxclsq (-p 0.1 –a), and a ML tree was inferred using RAxML v8.2.3 for each ortholog.

The species tree was inferred using the same two methods as above (i.e., using MQSST as implemented in ASTRAL-II and through a supermatrix ML analysis using FastTree to generate an input topology for a more thorough analysis using RAxML v.8.2.3. In both ML analyses, the WAG model of evolution was used, with partitioning by gene to ensure that a separate rate was estimated for each gene using CAT.

Nyctaginaceae—The node level clustering that contained the Nyctaginaceae (37 taxa) and Petiveriaceae (4 taxa) was used for inference of the clade (node 2; Fig 4-2). Initially, homolog groups, which were found to contain at least 1000 genes and sequences from both Nyctaginaceae and Petiveriaceae, were aligned using MAFFTv.7, cleaned with pxclsq for 10% matrix occupancy, and homolog trees were inferred with FastTree v2.1.8 under the WAG model of evolution. Next, spurious tips with a relative value of 1 substitution/site and an absolute value of 2 substitutions/site were removed and monophyletic tips were masked conserving the tip with the highest number of aligned characters. Next, orthologs were inferred using the Maximum-Inclusion procedure, searching for ortholog groups containing at least 40 of the 41 taxa, which resulted in 389 orthologs for the analysis. Species trees were inferred using the method mentioned above for Caryophyllaceae.

Cactaceae—The species trees were inferred using the same method as Nyctaginaceae with the following minor modifications. The cluster used was the node-level cluster that consisted of Cactaceae (29 taxa), Portulacaceae (8 taxa) and Anacampserotaceae (3 taxa) (node 6; Fig. 4-2). The ortholog groups were filtered for those consisting of at least 40 of the 47 taxa, which resulted in 1502 orthologs.

Amaranthaceae and Chenopodiaceae—The species trees were inferred using the same method as Nyctaginaceae with these minor modifications: we used homologous gene clusters of 500 sequences or fewer, as opposed to 1000. The clusters used were from the node-level cluster that consisted of Amaranthaceae (21 taxa), Chenopodiaceae (39 taxa), and five representative Caryophyllaceae (node 8; Fig. 4-2). The ortholog groups were filtered for those consisting of at least 60 of the 65 taxa, which resulted in 455 orthologs.

The non-core Caryophyllales—The species trees were inferred using the same method as Nyctaginaceae with the following modifications. The cluster used was the node-level cluster consisting of Polygonaceae (37 taxa), Plumbaginaceae (4), Tamaricaceae (3), Nepenthaceae (3), and Droseraceae (4) (node 12; Fig. 4-2). This was combined with the clustering of smaller families to add in Drosophyllaceae (1), Ancistrocladaceae (1), Frankeniaceae (2) and Basellaceae (2), Microteaceae (1), Physenaceae (1) and Simmondsiaceae (1) were added as outgroups. The ortholog groups were filtered for those consisting of at least 55 of the 60 taxa, which resulted in 514 orthologs, of which only 513 contained at least one outgroup and were rooted for the conflict analysis. The final statistics for the supermatrix can be found in Appendix C.

Analysis of conflict

We conducted conflict analyses on the trees resulting from the IND analyses using the bipartition-based method as implemented in phyparts (Smith et al., 2015). All gene trees from the clade-specific analyses were rooted by outgroups in a ranked fashion using pxrr (from the phyx package; Brown et al. 2017), whereby, if a taxon in the outgroup is not found the program searches for the next taxon, thus not requiring all outgroup taxa for rooting. The results were summarized and mapped onto a tree using phypartspiecharts.py (https://github.com/mossmatters/MJPythonNotebooks). A comparison of conflict between the topology of the MQSST and the ML analysis was conducted using pxbp (from the phyx package; Brown et al. 2017), where both trees were rooted on all outgroups using pxrr and the MQSST tree was mapped onto the ML tree.

Results

We define the support on the MQSST species tree from here on as follows: strong support will correspond to local posterior probabilities (LPP) ≥ 0.95, moderate support will correspond to 0.95 > LPP ≥ 0.80, and low support will correspond to 0.80 > LPP. For the bootstrap (BS) support on the ML tree we will consider strong support to be BS ≥ 90, moderate support will be 90 > BS ≥ 70, and poor support will be anything with BS support lower than 70. Here we also discuss the Quartet Differential (QD), which reflects the number of alternate topologies a quartet recovers. This method provides a means of disentangling a rogue node from one with two dominant topologies and a thorough description of this form of support and other quartet based support metrics is outlined by Pease et al. (2018).

We inferred species relationships using multiple datasets— one dataset comprised all taxa (ALL; Fig. 4-3) whereas the other datasets (described in Appendix C) included only orthologs inferred from five most densely sampled clades (IND; Figs. 4-4—4-8).

Relationship among major clades across the backbone of Caryophyllales using the ALL dataset

Both ML and MQSST analyses recovered a clade of Tamaricaceae+Frankeniaceae sister to Plumbaginaceae+Polygonaceae, which we will collectively refer to as the non-carnivorous non-core (NCNC) clade (Fig. 4-3). The MQSST analysis had insufficient data to resolve the divergence of the NCNC, resulting in no branch length at the divergence of the carnivorous clade (the families Droseraceae, Ancistrocladaceae, Drosophyllaceae, and Nepenthaceae), and the core Caryophyllales (all other families). In the ML analysis we recovered the carnivorous clade to be sister to the core Caryophyllales and the NCNC with low support from the ML support statistics. The majority of nodes within core Caryophyllales received medium to high support in the MQSST and ML trees with notable examples of low support occurring in Amaranthaceae subfamily Polycnemoideae (*Polycnemum majus* and *Nitrophila occidentalis*) and the placement of Cactaceae genera *Leuenbergeria* and *Pereskia*.

The core Caryophyllales was inferred to be nested within a grade of species-poor families (Fig. 4-3). In the MQSST tree this grade consisted of Simmondsiaceae, Physenaceae, Microteaceae, and a clade of Stegnospermataceae+Macarthuriaceae diverging in that respective order. In the ML analysis, Limeaceae is nested within the grade, diverging prior to Stegnospermataceae+Macarthuriaceae. The grade is strongly supported in the MQSST analysis, whereas in the ML analysis there is low bootstrap support for the position of Limeaceae, which in combination with a QD of 0.38 towards a different topology indicates it may have bias towards an alternate position than that recovered by the ML analysis.

Caryophyllaceae was inferred in both analyses to be sister to Achatocarpaceae+Amaranthaceae+Chenopodiaceae, with Amaranthaceae+Chenopodiaceae forming a clade sister to Achatocarpaceae. In the MQSST analysis, Chenopodiaceae was monophyletic; however, the subfamily Polycnemoideae was not nested within Amaranthaceae, making Amaranthaceae paraphyletic without the inclusion of Chenopodiaceae. In the ML analysis, there was low support for a clade consisting solely of genus *Beta* and Polycnemoideae, making Chenopodiaceae paraphyletic without Amaranthaceae. Sister to the clade containing Amaranthaceae and Chenopodiaceae, the family Achatocarpaceae was recovered as monophyletic with strong support by both BS and LPP and no common discordant topologies were found from the QS analysis.

Sister to the clade of Amaranthaceae and relatives was a clade encompassing the family Nyctaginaceae and relatives, which, in the MQSST analysis also contained Limeaceae. Both the ML and MQSST recovered a strongly supported clade that consisted of the families Kewaceae, Aizoaceae, Gisekiaceae, Sarcobataceae, Agdestidaceae, Phytolaccaceae, Petiveriaceae, and Nytaginaceae (Fig. 4-3). Kewaceae was sister to all others, with Aizoaceae diverging first amongst the remaining members, followed by the monotypic Gisekiaceae. The next lineage to diverge is a clade containing the family Phytolaccaceae as sister to a strongly supported clade including the families Sarcobataceae and Agdestidaceae. Next there is a strongly supported clade of Petiveriaceae+Nyctaginaceae.

The Portullugo clade containing Molluginaceae and Portulacineae was strongly supported as monophyletic. The family Molluginaceae was sister to the rest of the Portulacineae, with the divergences of Montiaceae, Basellaceae, Didieraceae and Talinaceae resolved as a grade (Fig. 4-3). This led to a clade in which the family Cactaceae was sister to a clade of Anacampserotaceae+Portulacaceae. The monophyly and placements of all families were strongly supported.

Phylogenetic resolution among and within major Caryophyllales families from IND analyses

Non-core Caryophyllales—The sampling of the non-core Caryophyllales consisted of 60 species, with at least one representative from eight of the nine families (Fig. 4-4). All species relationships from the IND analysis were congruent with those of the ALL MQSST with the exception of the placement of *Eriogonum longifolium*. The family Polygonaceae had the highest density of sampling with 37 taxa. The IND analysis of the ML and MQSST analyses had a final matrix occupancy of ~ 81% (Appendix C). The MQSST and the ML supermatrix analyses were largely congruent, aside from the genus *Eriogonum,* where all nodes contained at least 50% gene-tree discordance and a few relationships had low LPP support. The families of carnivorous taxa (including Ancistrocladaceae, which has reverted to be non-carnivorous) formed a clade.

The four families of the NCNC were also monophyletic—however, this was with medium LPP support and (>50%) gene-tree conflict.

Amaranthaceae/Chenopodiaceae—The results of the IND analysis MQSST and the ALL MQSST were concordant except in the position of the genus *Beta* and the species *Tidestromia lanuginosa*. Sampling consisted of 60 Amaranthaceae, 39 of the taxa were members of the former 'Chenopodiaceae' (Hernández-Ledesma et al., 2015; Byng et al., 2016), and five Caryophyllaceae samples were used as outgroups. The choice of orthologs used in the analysis contained at least 60 taxa, resulting in 455 orthologs with approximately 15.5% missing data in the supermatrix (Appendix C). The MQSST and ML analysis contained three discrepancies (Fig. 4-5), all of which were marked by a minimum of 75% gene tree discordance and non-perfect LPP support at the contentious node. The ML/MQSST conflict surrounded the relationships of the genus *Beta* where it is either sister to all other Chenopodiaceae or found nested within Chenopodiaceae. Another conflict was the relationship of *Krascheninnikovia lanata* and *Suckleya suckleyana,* where the two taxa appeared as sister in the supermatrix analysis, but showed *S. suckleyana* and *K. lanata* formed a grade in the MQSST analysis (Fig. 4-5).

The majority of missing sequence data for the analysis was found in the clade that consists of the genus *Suaeda,* and the position of *Bienertia* as sister to *Suaeda* had a dominant alternative topology that consisted of roughly the same number of gene trees as the rest of the conflict (Fig. 4-5). Most of the conflict in the relationships was located at deeper nodes along the phylogeny. We found 376 of the 455 gene trees conflicted with the species tree surrounding the paraphyly of Amaranthaceae, with *Nitrophila occidentalis* and *Polynemum majus* forming a clade sister to the species which were formerly recognized as Chenopodiaceae. Although gene tree concordance was low (~17%), there was no dominant alternative topology found among the conflicting topologies.

Cactaceae—The inferred topology from the ALL MQSST analysis was congruent with the IND MQSST analysis of Cactaceae, except for the relationship between the genera *Leuenbergeria* and *Pereskia* and the relationship of the genera *Gymnocalycium* and *Stetsonia* (Figs. 4-3&4-6). The Cactaceae sampling included 29 ingroup taxa and inference of the IND species tree was done using 1502 orthologs with ~19% missing data in the final supermatrix (Appendix C). The MQSST and ML supermatrix species trees contained a high level of gene-tree conflict among many relationships (Fig. 4-6). This included whether the non-succulent taxa, previously circumscribed as *Pereskia* (now *Leuenbergeria* and *Pereskia*), were monophyletic or paraphyletic. High gene-tree conflict (>75%) was prevalent across many relationships including the position of *Lophophora williamsii*, the relationship of *Salmiopuntia salmiana* and *Tunilla corrugata,* and the relationship of the genus *Pereskia* with respect to the genus *Leuenbergeria*. Most of the missing data for the analysis was from the two species in the genus *Pereskia*.

Caryophyllaceae—The topologies of the all-species MQSST analysis and the IND MQSST analysis were completely concordant. The sampling across Caryophyllaceae consisted of 39 ingroup taxa and inference of the IND species tree was done using 999 orthologs, with ~17% missing data in the final supermatrix (Appendix C). The MQSST and ML supermatrix species tree analyses resulted in congruent topologies, with perfect LPP support at almost all nodes (Fig. 4-7). Most genera were recovered as monophyletic, with the exception of *Arenaria,* where almost all gene trees placed *Arenaria procera* sister to *Eremogone hookeri.*

Nyctaginaceae—The ALL MQSST analysis was concordant with the IND analysis aside from the position of *Boerhavia ciliata.* The sampling across the Nyctaginaceae consisted of 37 Nyctaginaceae with 4 Petiveriaceae used as outgroups. The species tree was inferred using 389 orthologs with ~14% missing data for the final ML supermatrix. The MQSST and supermatrix IND analyses were largely congruent aside from the relationships among species in the genus *Boerhavia* (Fig. 4-8). Within *Boerhavia*, there were 111 gene trees supporting *Boerhavia coccinea* sister to *Boerhavia torreyana* and 107 gene trees supporting an alternative of *B. coccinea* sister to *Boerhavia purpurascens*. The incongruent node contains a large amount of conflicting gene-tree signal with a dominant alternative topology matching the MQSST analysis. The node supporting the monophyletic herbaceous xerophytic clade contains almost no gene-tree conflict.

Discussion

Utilizing broad and dense transcriptome sampling for inference in Caryophyllales

Previous phylogenetic analyses have vastly improved our understanding of the backbone relationships of Caryophyllales (Rodman et al., 1984; Cuenod et al., 2002; Brockington et al., 2009; Schäferhoff et al., 2009; Yang et al., 2015), but strong resolution of early diverging lineages has proven a formidable task. Here, with increased taxon sampling and larger datasets, we reconstructed most relationships with high support (Fig. 4-3). This was true for deeper-level relationships that previously had weak or moderate support (e.g., Sarcobataceae and Agdestidaceae), as well as for new hypotheses (e.g., Stegnospermataceae as sister to Macarthuriaceae). Reassuringly, and similar to the results other phylogenomic studies (Cannon et al., 2015; Yang et al., 2015; Pease et al., 2016), we find most relationships in the tree are concordant with previous single- or multi-gene studies. This indicates that, in many cases, data sets of one or a few genes are sufficiently powerful for inferring most species relationships. While this improved resolution highlights the power of large nuclear datasets for phylogenetic inference, it is important to note that such data sets are not a phylogenetic panacea. For example, our analyses conflicted with the previously inferred monophyly of the families within the non-core Caryophyllales. Both MQSST and ML analyses found non-core Caryophyllales to be non-monophyletic (which was weakly supported as monophyletic in Yang et al. (2015)), the MQSST placed the carnivorous Caryophyllales with 0 branch length and no support as sister to the core Caryophyllales. The ML analyses weakly supported the non-carnivorous non-core as sister to the core Caryophyllales.

The inability of >1000 orthologs to provide statistical support for this relationship demonstrates a limitation of the current methods for phylogenetic inference with large datasets. This may be due to methodological limitations (e.g., model misspecification or oversimplification) or biological reality (e.g., biological processes occurred that obfuscate this relationship and leave little to no informative signal). Many relationships are the result of complex evolutionary histories that are manifested in conflict among gene tree topologies. Although conflict makes it difficult to infer species relationships, phylotranscriptomics provides a cost-efficient means of identifying conflicting gene trees and hence potentially exposing the underlying evolutionary processes, including ILS, hybridization, and gene duplication, that are often masked when using a small number of genes. Some of these recalcitrant phylogenetic relationships may be resolved by more sophisticated methods (Olave et al., 2015) but some may never be resolved due to the complex nature of evolution and speciation (e.g. hybridization, ILS, and gene duplication and loss). This can even lead to cases of “hard polytomies” originating when lineages radiate almost simultaneously from a common ancestor.

The analyses presented here add to a growing number of phylogenomic analyses that have exposed extensive underlying gene-tree conflict (Smith et al., 2015; Pease et al., 2016; Walker et al., 2017). Methods for analyzing and incorporating this conflict are rapidly emerging (Ané et al., 2007; Leigh et al., 2008; Salichos et al., 2014; Smith et al., 2015; Kobert et al., 2016; Arcila et al., 2017). We found, as with previous studies, that gene-tree conflict was unevenly distributed. For example, clades that may have undergone a rapid radiation (e.g., Cactaceae) (Arakaki et al., 2011) exhibit more gene-tree conflict than others (e.g., Caryophyllaceae). In some cases, we found nodes with as few as 50 out of 455 gene trees (~17%) supporting the ML and MQSST relationship (e.g., the position of the subfamily Polycnemoideae within the Amaranthaceae/Chenopodiaceae). However, in this case the relationship with the next most gene tree support 29 out of 455 (~6%) recovered the Polycnemoideae as sister to both Chenopodiaceae/Amaranthaceae.

Several instances of gene-tree conflict may have important taxonomic implications—for example, the most commonly inferred relationship from our molecular data indicate Polycnemoideae are more closely related to Chenopodiaceae, while they are morphologically more similar to Amaranthaceae and group with Amaranthaceae s.s. in molecular studies based on chloroplast gene regions (Masson & Kadereit, 2013 and ref. therein). Many traits in Polycnemoideae appear plesiomorphic and may have resulted from hybridization or ancestral polymorphism. Regardless of the underlying reasons, identifying relationships with high gene-tree conflict illustrates the power of large datasets to document evolutionary processes that cannot be elucidated with phylogenies containing only a few genes. Development of new methodologies for identifying and analyzing gene-tree conflict is an essential step forward for understanding species relationships.

Evaluating the patterns and causes of gene tree conflict results in a more informed and nuanced understanding of evolutionary history. For example, the earliest branches within the former genus *Pereskia* s.l. (now split into *Pereskia* and *Leuenbergeria*) displayed high levels of gene-tree conflict. Both *Pereskia* and *Leuenbergeria* share many defining morphological features, however, the species tree inference resolved them to form a grade as previously demonstrated by Edwards et al. (2005). As molecular studies comprehensively examine genera and families, we may begin to better understand why some morphological features fail to match molecular phylogenies. In a broader sense, using phylogenomic datasets to understand the complex processes that may hide beneath perfect bootstrap support will add greater depth to the field of systematics, elucidating the complexities of the evolutionary processes responsible for adaptations that have shaped the world around us.

Taxonomic results for Caryophyllales

Agdestidaceae—Our analyses strongly support the sister relationship of *Sarcobatus* and *Agdesits* suggested in several previous analyses, these typically with weak to moderate support (Brockington et al., 2011; Cuénoud et al., 2002; Schäferhoff et al., 2009). Given this relationship, and given the significant differences in floral morphology, habit, wood anatomy, etc. that characterize these genera, we suggest that both be treated as monogeneric families, as advocated by Hernández-Ledesma et al. (2015).

Amaranthaceae s.l.—The monophyly of the traditional Chenopodiaceae in our analyses, including its sister relationship to subfamily Polycnemoideae builds upon a growing body of evidence that suggests the broad circumscription of Amaranthaceae *sensu* APG (2016) may need to be reevaluated. Polycnemoideae is disjunctly distributed in Eurasia, America and Australia and consists of only 13 (mostly rare) species in four genera which considering the Eocene stem age appears as a relictual lineage (Masson and Kadereit, 2013). Molecular phylogenetic studies based on chloroplast markers and extensive sampling (Kadereit et al., 2003, 2012) as well as morphological similarities (petaloid tepals, filament tubes, 2-locular anthers; compare Kadereit et al., 2003: Tab. 5) place them closer to the Amaranthaceae s.str., while in terms of habitat preferences they are more like many members of the Chenopodiaceae. Our analysis contradicts the placement of Polycnemoideae in Amaranthaceae s.str. as proposed by Masson & Kadereit (2013) and provides evidence that it forms a clade sister to Chenopodiaceae in 17% of the gene trees. Nevertheless, some key early-diverging lineages in the Amaranthaceae s.l. clade are missing from our analyses (e.g*., Bosea* and *Charpentiera*), and hence additional taxon sampling will be necessary to address these contradictory results.

Future directions for phylogenomic analyses of the Caryophyllales

Although we found strong resolution for many relationships among the Caryophyllales, our analysis highlights several key nodes with weak support that would benefit from more focused analyses. These include additional sampling of the missing Caryophyllales families as well as expanded sampling within major subclades of the order. For example, our results highlight the need for future investigation into the non-core Caryophyllales to explore the conflict at deep nodes in this area of the tree. The group has previously been recognized or treated as monophyletic (Brockington et al., 2009, 2011; Walker et al., 2017; Yang et al., 2017), and the poor resolution in our analyses hampers our understanding of key evolutionary events in this group (e.g., evolution of endosperm, production of secondary compounds, evolution of plant carnivory).

More extensive sampling within several families may also be necessary to resolve relationships and explore gene tree conflict in several other areas of Caryophyllales phylogeny. For example, the discrepancy between the MQSST and the ML analyses in the placement of the family Limeaceae may be affected by the inclusion of only one species of *Limeum*. However, a phylogenetic study with greater taxon sampling of *Limeum* (Christin et al., 2011) agreed with the MQSST topology presented here. In any case, it is important to resolve the position of Limeaceae given its importance to the understanding of the complex pigmentation patterns seen in core Caryophyllales (Brockington et al., 2015; Lopez-Nieves et al., 2017). Further studies of Molluginaceae would also be valuable for their insight into C_4 evolution, as would more targeted studies of its sister clade the Portulacineae. More specific analyses using transcriptome data have helped uncover adaptive gene family expansions in Portulacineae (Wang et al., 2018), multiple paleopolyploidy events in the carnivorous Caryophyllales (Walker et al., 2017), and are warranted to explore the convergent evolution of the many other extreme adaptations across Caryophyllales. Some of these include the evolution of cold tolerance across Caryophyllaceae and Polygonaceae, multiple origins of C_4 photosynthesis in Amaranthaceae s.l., and the evolution of drought tolerance in Nyctaginaceae, Polygonaceae, Aizoaceae, and Portulacineae.

Future directions for large-scale phylogenomic studies

Transcriptomics has emerged as a powerful tool for phylogenomics. The ever-decreasing costs of sequencing combined with improved methods for collecting plant material (Yang et al., 2017) and downstream data analysis (Dunn et al., 2013; Kocot et al., 2013; Yang and Smith, 2014; Emms and Kelly, 2015; Washburn et al., 2017) have made this a cost-efficient means for investigating systematic and evolutionary questions. To date, phylotranscriptomic analyses have been used at multiple phylogenetic levels, from genera (Pease et al., 2016; Yu et al., 2017) and large clades (Yang et al., 2015; McKain et al., 2016; Yang et al., 2017), to across all land plants (Wickett et al., 2014). As the size of these analyses continues to expand, so does their computational burden—a problem of critical importance for future research. This has never been more relevant for the botanical world than it is now, with the anticipated sequencing of 10,000 plant genomes (doi:10.1126/science.aan7165).

One challenge to increasing the size of phylogenomic datasets is the burden of homology identification. Here we explored a new approach for that attempts to divide and conquer the daunting task of homology identification, breaking with the typical all-by-all BLAST procedure. By dividing the transcriptomes into smaller homology problems before combining homolog groups with a post-order (tip-to-root) method, based upon a previously hypothesized phylogeny, we dramatically reduced one major computational burden (i.e., the scaling an all-by-all BLAST). Additionally accurate orthology detection is a key component of phylogenomic analyses, as demonstrated by a recent study demonstrating that two misidentified orthologs altered the species tree topology in a >200 gene dataset (Brown and Thomson, 2017). And so, this procedure also incorporated phylogenetic estimation into orthology detection (Gabaldón, 2008; Yang and Smith 2014; Yang et al. 2015) as BLAST is not a phylogenetically informed means of inferring relationships (Smith and Pease, 2016).

This hierarchal method of homology identification relies on some previously identified phylogenetic relationships. After clustering individual clades, each set of clusters is then combined (moving from tips to root) in an order defined by a simplified phylogeny. There are some benefits to this approach as, for example, it factors in clade-specific evolutionary history (e.g. a shift in molecular rate introduced from transition from a woody to an herbaceous life history). However, it may also introduce some bias as 1) it relies on some simplified phylogenetic relationships deep in the tree and 2) it assumes that the clustered groups form clades. If the groups clustered toward tips do not form a clade, clusters may be artificially broken up due to increased molecular distance of the included samples (i.e., distant species compared). While this may result in fewer homologs, this scenario is not likely alter an inferred species topology. For example, in our analyses, the phylogenies recovered Polycnemoideae as sister to Chenopodiaceae. However, during homology inference, Polycnemoideae was *a priori* clustered with Amaranthaceae. The clustering did not force Polycnemoideae to be sister to Amaranthaceae, but clustering Polycnemoideae with Chenopodiaceae first may have resulted in more recovered homologs. However, as with any method further investigation is warranted.

We found the tip-to-root method to be a powerful means of reducing the computational time spent conducting an all-by-all BLAST across the entire dataset. However, clustering analyses that involve deep splits in the angiosperm tree of life tend to result in reduced dataset size in terms of number of useful orthologs and homologs. For example, a comparison of the number of identified orthologs between the tip-to-root clustering in the current study and an all-by-all BLAST of the non-core Caryophyllales study of Walker et al., 2017, (that included 10 ingroup and two outgroup taxa) showed a greater number of inferred orthologs from the latter dataset. Walker et al. (2017) recovered 1637 orthologs high matrix occupancy (i.e., most or all orthologs present for all taxa), whereas in the current study, 514 orthologs were recovered with high matrix occupancy. This discrepancy may be due to homologs being filtered by one of several cutoffs or systematic error due the difficulty of inferring larger homolog phylogenies.

This presents an interesting dichotomy. Increasing taxon sampling and phylogenetic breath and depth can improve accuracy and alter the inferred relationships and support. However, increased taxon sampling greatly increases the complexity and burden on each step in the inference process. Further explorations and methods will be required to fully realize the potential of these datasets and allow for their continued growth.

Homology detection and gene-tree conflict are not the only analytical burdens that future phylogenomic studies should seek to improve. Additional computational complexities such as evolutionary rate heterogeneity, distinguishing between ILS and hybridization, and improved understanding of gene duplication and loss will be important considerations for improving future phylogenomic analysis. While these are beyond the scope of this paper, the continued growth of phylogenomics portends an exciting time of evolutionary discovery.

Acknowledgements

We thank G. Stull and M.J. Grangé-Guermente for helpful comments on preparation of the manuscript; A. Summers, N. Douglas, A. Solís Magallanes, A. Clifford, S. Lavergne, T. Stoughton, N. Jensen, W. Judd, D. Warmington, S. Pedersen, K. Thiele, and J. Brittnacher for assisting with obtaining plant materials; the Cambridge University Botanic Gardens, Bureau of Land Management, US Forest Service, the Navajo Nation, California State Parks, Missouri Botanical Garden, Rancho Santa Ana Botanic Garden, Desert Botanical Garden, the Kampong of the National Tropical Botanical Garden, Sukkulenten-Sammlung Zürich, Cairns Botanic Gardens, Botanischen Gartens--Technische Universität Dresden, Millennium Seed Bank, and Booderee National Park for granting us access to their plant materials; M.R.M. Marchán-Rivadeneira, L. Cortés Ortiz, V. S. Mandala, R. Mostow, M. Croley, L. Leatherman, and M. Parks for help with lab work, and M. Palmer at Matthaei Botanical Gardens for help with growing plants. We also thank the Associate Editor and two anonymous reviewers for comments that helped improve the manuscript. The molecular work of this study was conducted in part in the Genomic Diversity Laboratory at the University of Michigan.

Funding

Support came from University of Michigan, Oberlin College, the National Geographic Society, a Rackham Pre-doctoral Fellowship, a Nancy W. Walls award, and the National Science Foundation (DEB 1054539, DEB 1352907 and DEB 1354048).

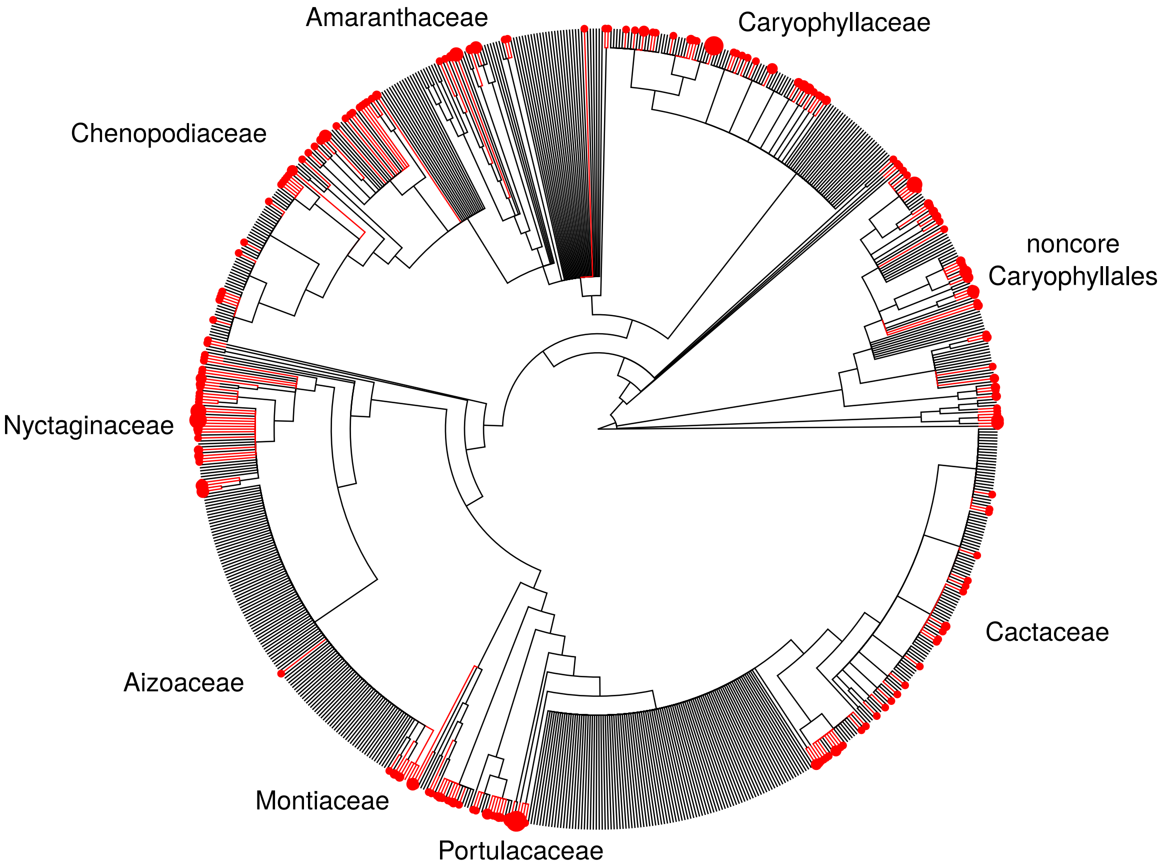


Figure 4-1 Transcriptome sampling across the Caryophyllales

Species level tree of the Caryophyllales from Smith et al. 2018, collapsed to genus level. Branches of genera included in the study are highlighted in red and circles at the tips are proportionate to the number of samples from a given genus.

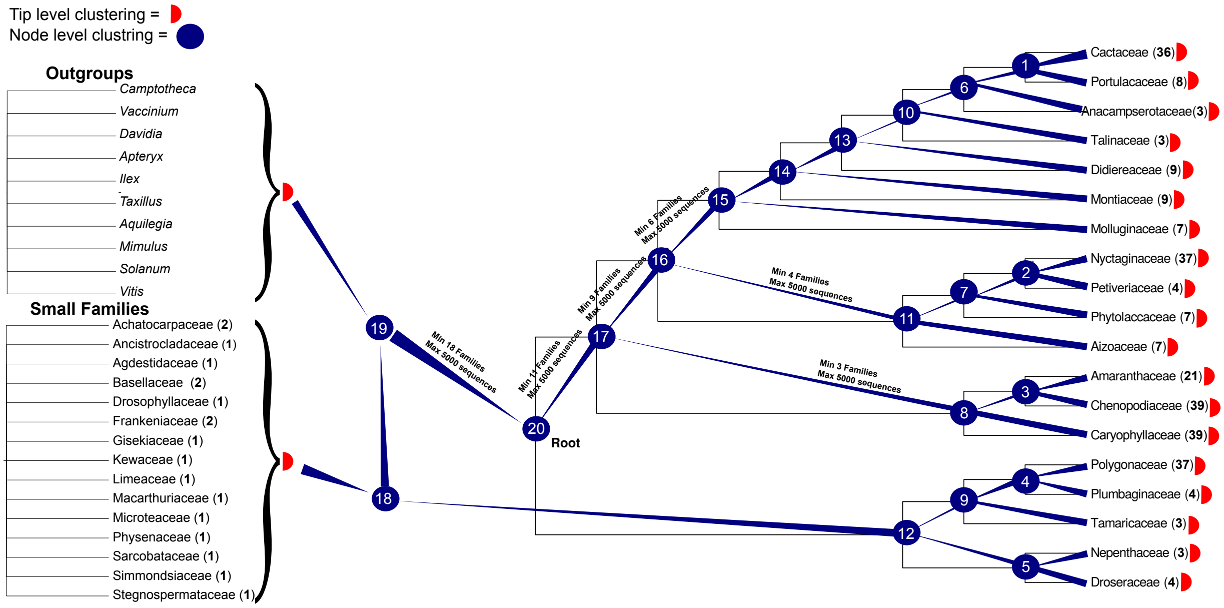


Figure 4-2 Representation of the post-order clustering method

Diagram of the general order in which clustering was performed, based upon a synthesis of previous phylogenies (Stevens, 2015). For outgroups, the name of the genus was given and, for Caryophyllales, the family name. In brackets, the number of individuals from that family sequenced is listed. Semi-circles represent tip-level all-by-all clustering and full circles represent node level clustering, with numbers representing order of clustering.

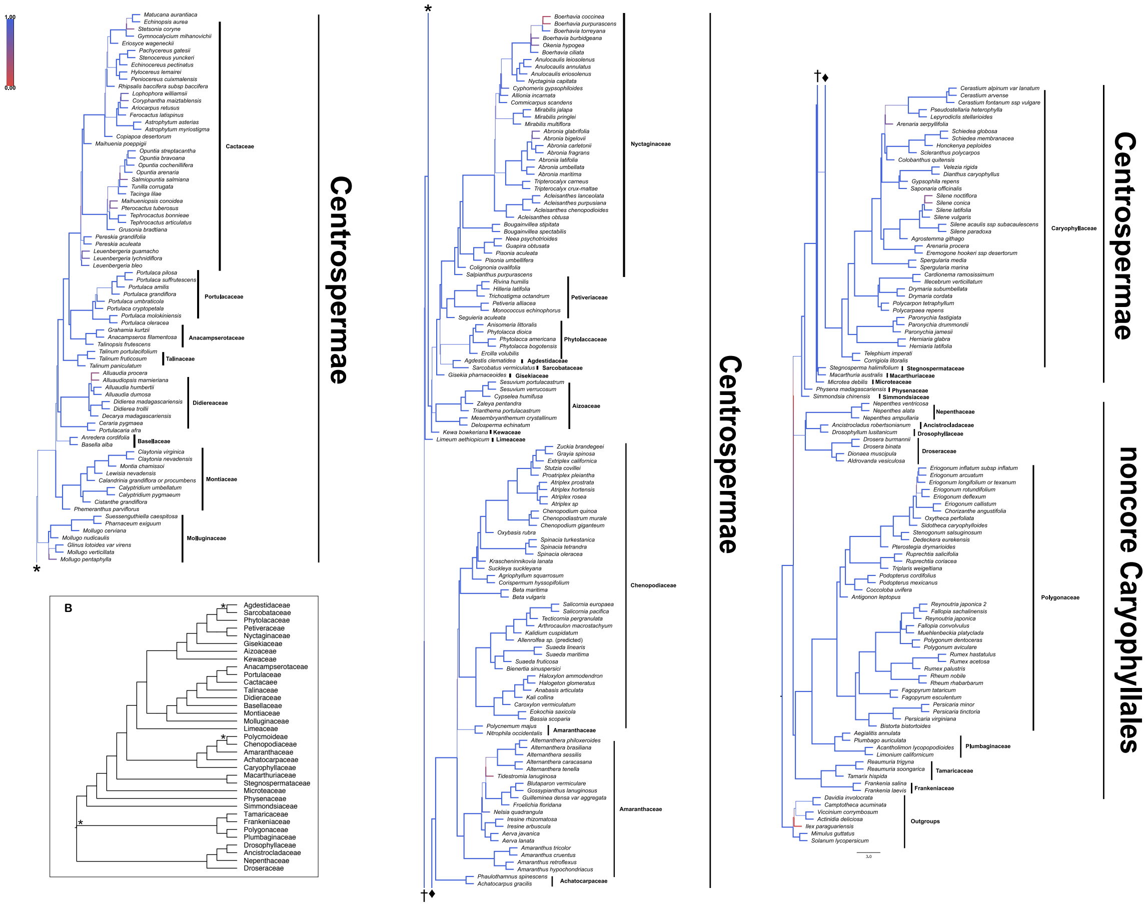


Figure 4-3 Caryophyllales phylogeny inferred from 305 transcriptomes

The Maximum Quartet Support Species Tree inferred from 305 transcriptomes. Branches are colored in a gradient to represent support, with cooler colors (Blue) representing strong support and warmer colors (Red) representing weak support. B) A tree showing the relationships among major families, with stars depicting major family level findings.

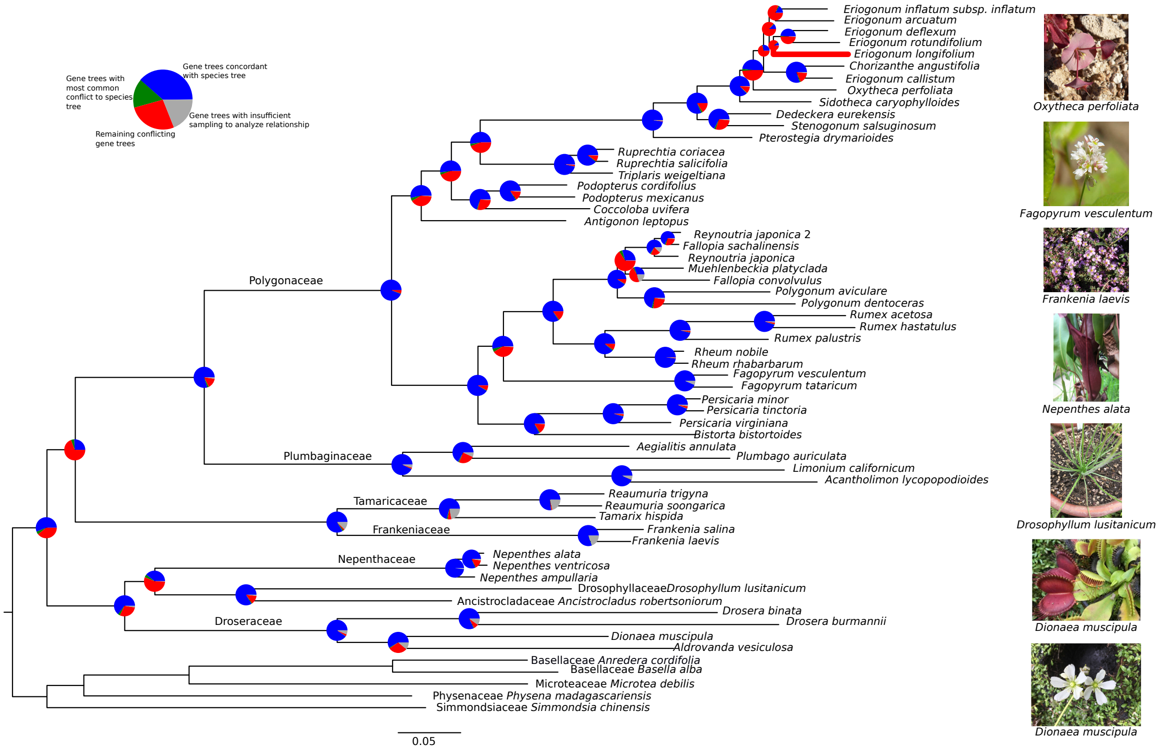


Figure 4-4 Inferred species relationships among taxa in the families of the non-core Caryophyllales

Phylogeny inferred using maximum likelihood (ML) from the concatenated dataset of the 514 inferred orthologs across the non-core Caryophyllales. Branches in red represent conflict with the maximum quartet support species tree. Gene tree conflict is represented as pie charts on the ML tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. Photo credits: *Oxytheca porfoliata*—Stan Shebs, *Fagopyrum vesculentum*—Kurt Stüber, *Frankenia laevis*—Ghislain118, *Nepenthes alata*—Joe Walker, *Drosophyllum lusitanicum*—Joe Walker, *Dionaea muscipula* (trap and flower)— Joe Walker. Licenses and location of original photographs can be found in Appendix C.

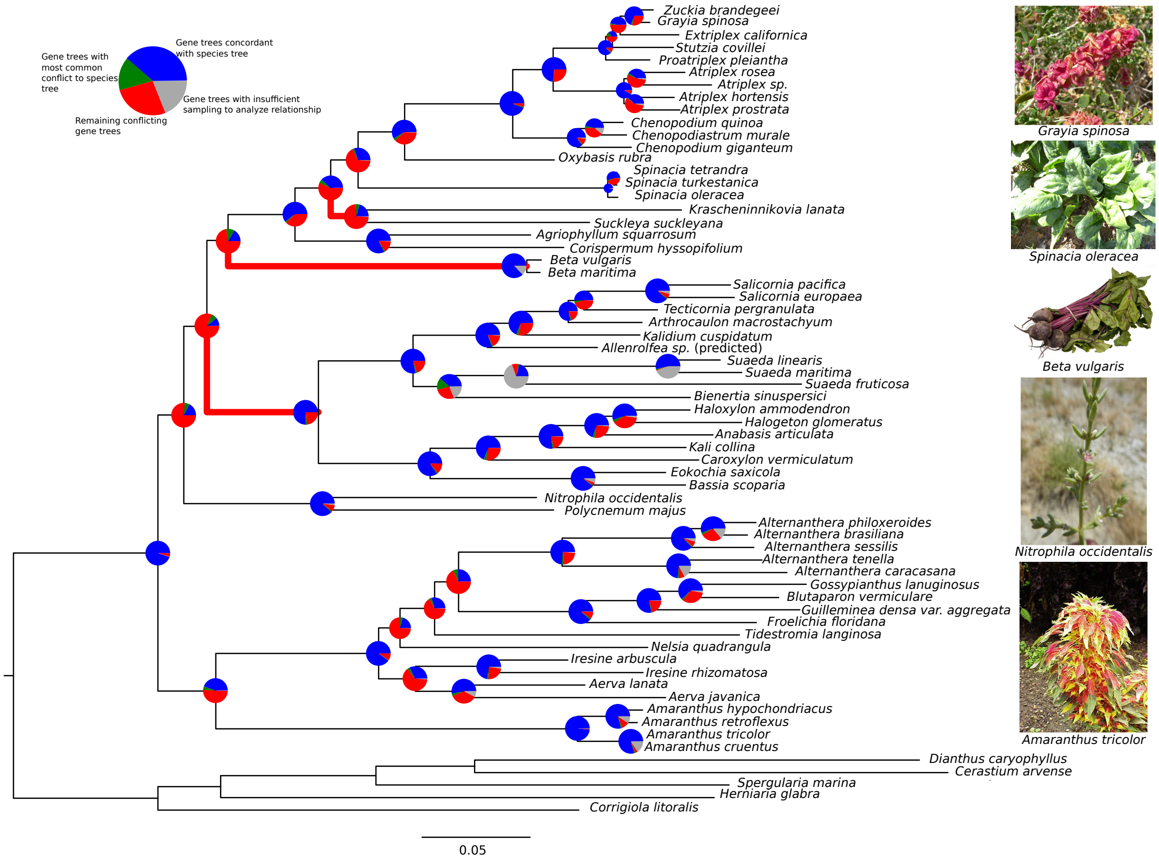


Figure 4-5 Inferred species relationships among taxa in the families Amaranthaceae and Chenopodiaceae

Phylogeny inferred using maximum likelihood (ML) from the concatenated dataset of the 455 inferred orthologs across the Amaranthaceae and Chenopodiaceae. Branches in red represent conflict with the maximum quartet support species tree. Gene tree conflict is represented as pie charts on the ML tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. Photo credits: *Grayia spinosa*—Stan Shebs, *Spinacia oleraceae*—Victor M. Vincent Selvas, *Beta vulgaris*—Evan Amos, *Nitrophila occidentalis*—Mike Moore, *Amaranthus tricolor*-- Kurt Stueber. Licenses and location of original photographs can be found in Appendix C.

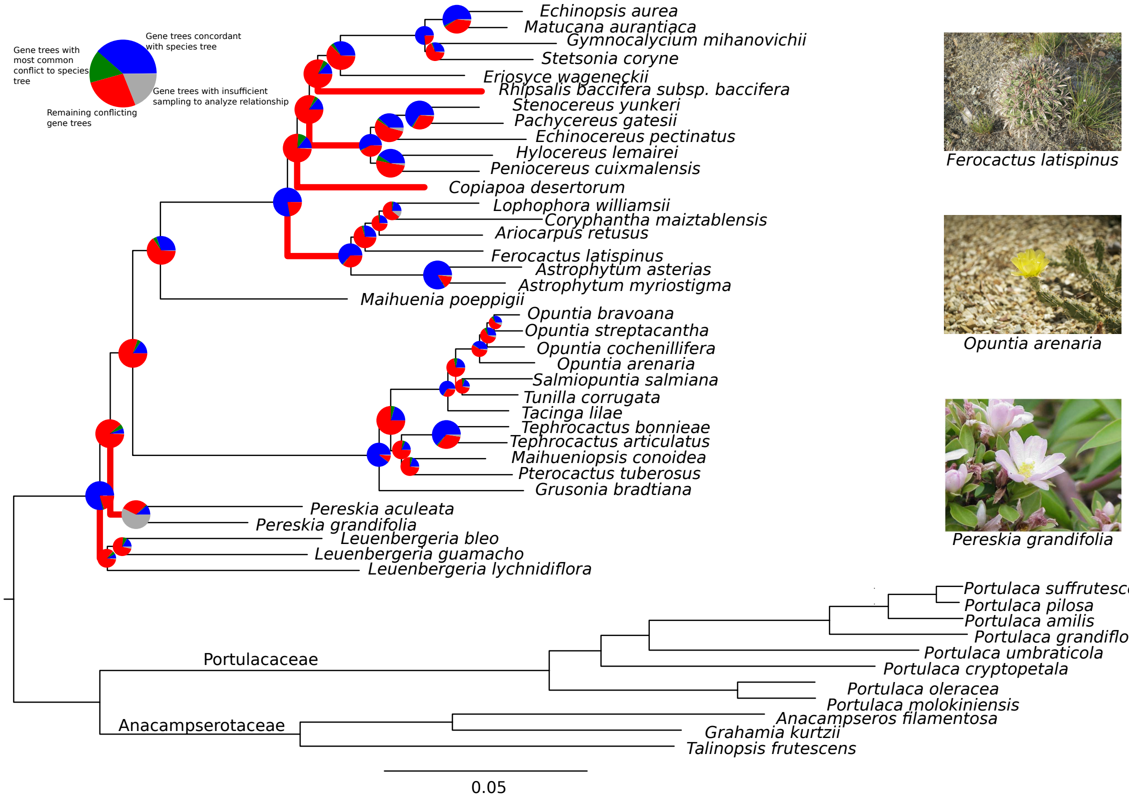


Figure 4-6 Inferred species relationships among taxa in the Cactaceae

Phylogeny inferred using maximum likelihood (ML) from the concatenated dataset of the 1502 inferred orthologs across the Cactaceae. Branches in red represent conflict with the maximum quartet support species tree. Gene tree conflict is represented as pie charts on the ML tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. Photo credits: *Ferocactus latispinus*—Lucas C. Majure, *Opuntia arenaria*—Lucas C. Majure, *Pereskia grandiflora*—Kurt Stüber. Licenses and location of original photographs can be found in Appendix C.

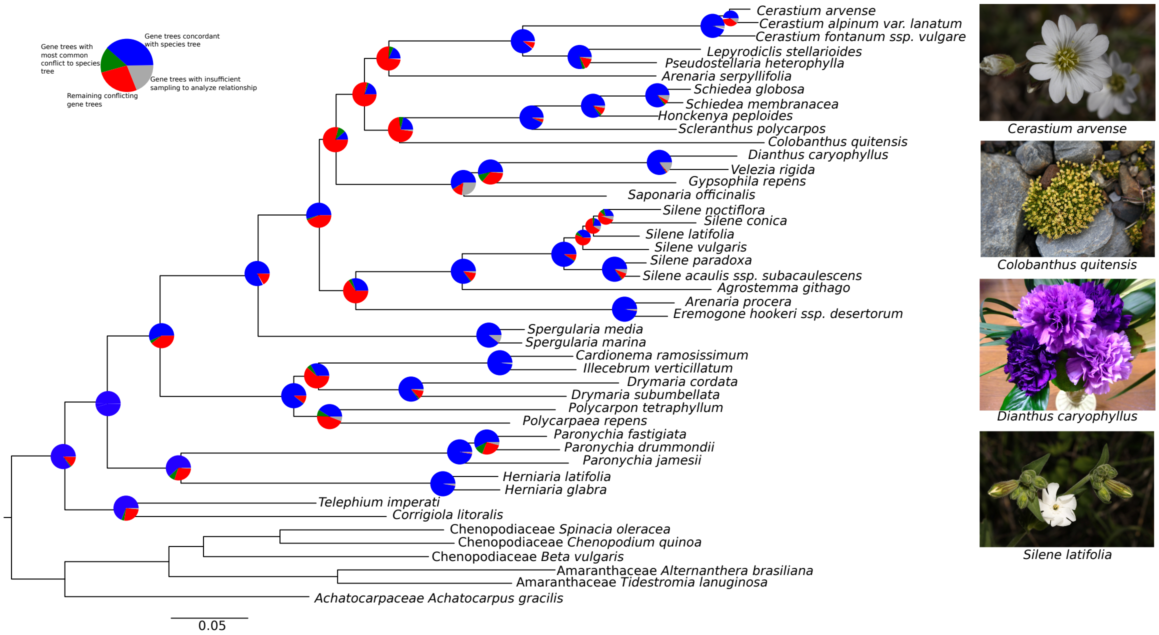


Figure 4-7 Inferred species relationships among taxa in the Caryophyllaceae

Phylogeny inferred using maximum likelihood (ML) from the concatenated dataset of the 999 inferred orthologs across the Caryophyllaceae. Gene tree conflict is represented as pie charts on the ML tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. Photo credits: *Cerastium arvense*—Walter Siegmund, *Colobanthus quitensis*—Liam Quinn, *Dianthus caryophyllus*—Pagemoral and *Silene latifolia*—Walter Siegmund. Licenses and location of original photographs can be found in Appendix C.

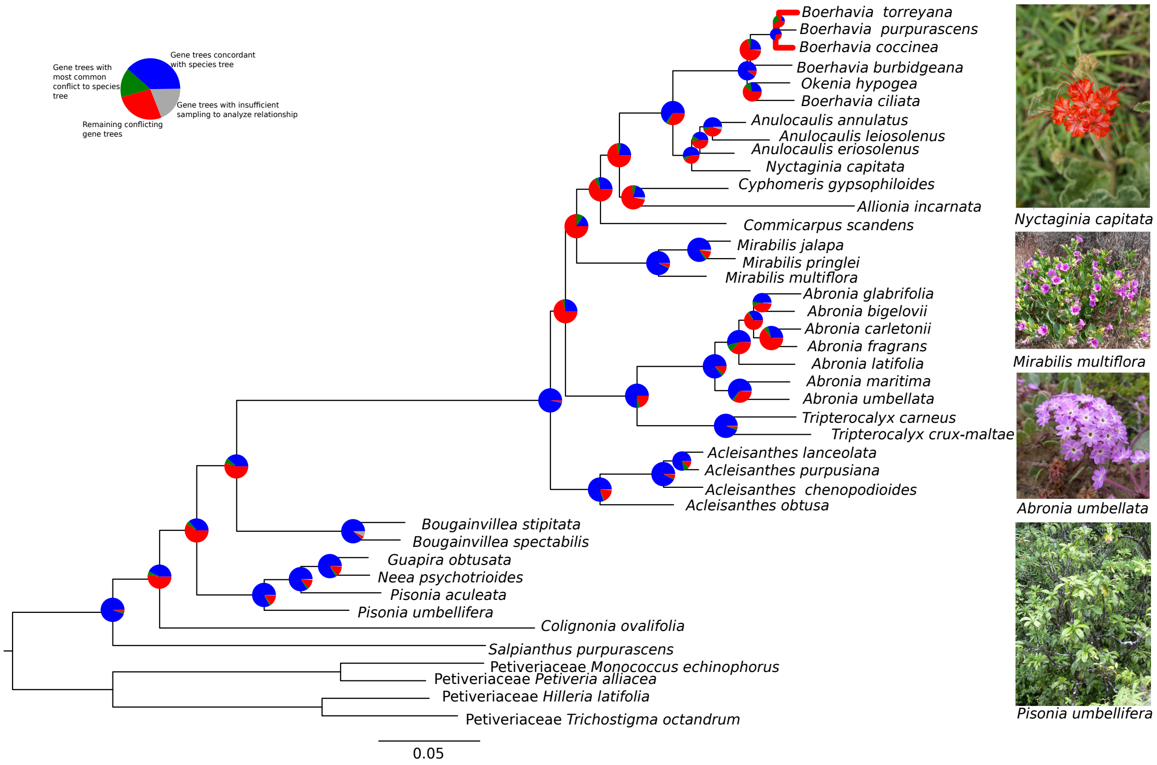


Figure 4-8. Inferred species relationships among taxa in the Nyctaginaceae.

Phylogeny inferred using maximum likelihood (ML) from the concatenated dataset of the 389 inferred orthologs across the Nyctaginaceae. Branches in red represent conflict with the maximum quartet support species tree. Gene tree conflict is represented as pie charts on the ML tree, blue indicates proportion of gene trees concordant with the ML tree topology, green indicates the most common alternative gene tree topology, red indicates conflicting gene trees with other alternative topologies, and grey indicates sampling was missing for the gene tree to infer a given relationship. Photo credits: *Nyctaginia capitata*— Mike Moore, *Mirabilis multiflora*—Mike Moore, *Abronia umbellata*—Mike Moore, and *Pisonia umbellifera*—Forest & Kim Starr. Licenses and location of original photographs can be found in Appendix C.

CHAPTER V

Analyzing contentious relationships and outlier genes in phylogenomics

Preamble: This chapter is our manuscript that is in press, the citation for this manuscript is: JF Walker, JW Brown and SA Smith. Analyzing contentious relationships and outlier genes in phylogenomics. In Press. ***Systematic Biology***.

Abstract

Recent studies have demonstrated that conflict is common among gene trees in phylogenomic studies, and that less than one percent of genes may ultimately drive species tree inference in supermatrix analyses. Here, we examined two datasets where supermatrix and coalescent-based species trees conflict. We identified two highly influential "outlier" genes in each dataset. When removed from each dataset, the inferred supermatrix trees matched the topologies obtained from coalescent analyses. We also demonstrate that, while the outlier genes in the vertebrate dataset have been shown in a previous study to be the result of errors in orthology detection, the outlier genes from a plant dataset did not exhibit any obvious systematic error and therefore may be the result of some biological process yet to be determined. While topological comparisons among a small set of alternate topologies can be helpful in discovering outlier genes, they can be limited in several ways, such as assuming all genes share the same topology. Coalescent species tree methods relax this assumption but do not explicitly facilitate the examination of specific edges. Coalescent methods often also assume that conflict is the result of incomplete lineage sorting (ILS). Here we explored a framework that allows for quickly examining alternative edges and support for large phylogenomic datasets that does not assume a single topology for all genes. For both datasets, these analyses provided detailed results confirming the support for coalescent-based topologies. This framework suggests that we can improve our understanding of the underlying signal in phylogenomic datasets by asking more targeted edge-based questions.

Introduction

Recent phylogenomic studies have shown that small changes to a dataset or the methods used to analyze a dataset can yield conflicting hypotheses at particular recalcitrant relationships with high support (i.e., 100% support from nonparametric bootstrap (BS) or posterior probability (PP) values). Prominent examples of this include many charismatic lineages such as the root of placental mammals (Morgan et al. 2013; Romiguier et al. 2013), early branching within Neoaves (Jarvis et al. 2014; Prum et al. 2015), and the earliest diverging lineage of extant angiosperms (Zanis et al. 2002; Wickett et al. 2014; Xi et al. 2014). The resolution of these relationships is critical to understanding the evolutionary history of their respective clades (e.g., patterns of biochemical, morphological, and life history evolution).

Finding the underlying causes of uncertainty in phylogenetic datasets is an essential step toward resolving problematic relationships. Recently, authors have developed means of exploring conflict between gene trees and species trees specifically for phylogenomic datasets (Salichos et al. 2014; Smith et al. 2015; Kobert et al. 2016), aiding in the identification of regions of species trees with considerable uncertainty despite strong statistical support from traditional support measures. Two studies have shown that the disproportionate influence of just one or two "outlier genes" on a supermatrix analysis is capable of driving tree topology inference (Brown and Thomson 2017; Shen et al. 2017). Using a Bayes factor approach Brown and Thomson (2017) reanalyzed a series of published datasets and found that the transcriptome data from Chiari et al. (2012) contained outlier genes. When outlier genes were included in phylogenetic reconstruction, a clade of turtles+crocodilians was inferred to be sister to birds with 100% PP. The same topology was previously inferred using ML with nucleotide data in the original study by Chiari et al. (2012) but was dismissed in favor of a coalescent reconstruction that placed turtles sister to birds+crocodilians. When Brown and Thomson (2017) removed the outlier genes, the reduced supermatrix inferred the same topology as the coalescent reconstruction with 100% PP. Another recently published study compared gene-wise likelihoods across multiple topologies to examine contentious relationships across the tree of life and found disproportionate influence of genes at all contentious relationships examined (Shen et al. 2017).

While such studies have highlighted several issues concerning phylogenomic conflict within datasets, these are early steps and several of these approaches have limitations that may limit our ability to identify phylogenetic support for particular relationships. For example, some of these analyses may incur significant runtimes that may limit more extensive dataset exploration or be a barrier for larger datasets. Also, these analyses are often performed on a small number (e.g., ~2) of alternative topologies (e.g., Castoe et al. 2009; Smith et al. 2011; Shen et al. 2017), and like typical supermatrix analyses, most explicitly assume that all genes share a topology. However, given widespread gene tree discordance (e.g., due to incomplete lineage sorting [ILS] and other processes), it may be more realistic to assume that many alternative topologies are supported within datasets (e.g., Smith et al. 2015; Pease et al. 2016; Walker et al. 2017). Coalescent species tree methods relax this assumption but typically assume that gene tree discordance is the result of ILS (but see Boussau et al. 2013). The computational burden of large datasets also typically limits these coalescent analyses to Maximum Quartet Support Species Tree (MQSST) methods (Mirarab and Warnow 2015) that have additional simplifying assumptions.

If the research question involves a small number of relationship and not the entirety of the tree, it may be more appropriate to examine targeted edges instead of resolved topologies (Lee and Hugall 2003). Here, we describe a fast analysis framework, maximum gene-wise edge (MGWE) analysis. This framework facilitates the examination of contentious edges in phylogenomic datasets without the requirement that each gene share the same topological resolution. We compare results from two-topology gene-wise log-likelihood and MGWE analyses for vertebrate (Chiari et al. 2012; Brown and Thomson 2017) and carnivorous Caryophyllales datasets (Walker et al. 2017) (hereafter referred to as the carnivory dataset). Both datasets contain contentious relationships, outlier genes, and, in their respective original studies, the authors dismissed the supermatrix topology for the topology inferred using a coalescent method. In both cases, we find that the use of an edge based approach results in stronger support for the topology hypothesized to be correct by researchers in the original study.

Materials and Methods

Data collection

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

Species trees

Brown and Thomson (2017) used Bayesian analyses to obtain the topologies from the Chiari et al. (2012) data set. As our study focused on the use of maximum likelihood (ML) for detecting overly influential genes, we ensured that ML phylogenetic reconstruction would recapitulate the previous species tree results. To construct a supermatrix tree for the vertebrate dataset, the 248 individual vertebrate genes used in Brown and Thomson (2017) were concatenated using the Phyx program pxcat (Brown et al. 2017). The species tree was inferred in RAxML v8.2.10 (Stamatakis 2014) using the GTR+ Γ model of evolution, and edge support was assessed from 200 rapid bootstrap replicates. Supermatrix trees for the vertebrate dataset were inferred both with all genes present, and again with the previously identified two outlier genes (8916 and 11434) removed (see below). The ML tree inferred from all the data from the carnivory dataset was downloaded from (http://dx.doi.org/10.5061/dryad.vn730http://dx.doi.org/10.5061/dryad.33m48) while a novel ML tree was inferred from a reduced supermatrix that excluded two highly influential genes (cluster575 and cluster3300; see below).

Gene tree construction and analysis of conflict

Individual gene trees for both datasets were inferred using ML with the GTR+ Γ model of evolution as implemented in RAxML. SH-like analyses (Anisimova et al. 2011), as implemented in RAxML, were performed to assess gene tree edge support. As this analysis examines alternative topologies by nearest-neighbor interchange (NNI), it is possible that during the analysis a topology with a higher likelihood is found (i.e., an 'NNI-optimal' topology). When a better topology was found, that topology was used in downstream analyses. We used the pxrr program in the Phyx package (Brown et al. 2017) to root all gene trees on the outgroup (*Protopterus* for the vertebrate dataset, and *Beta vulgaris* and *Spinacia oleraceae* for the carnivory dataset) and we excluded gene trees where an outgroup was not present. We mapped conflict onto the supermatrix tree using phyparts (Smith et al. 2015) with SH-like support of < 80 treated as uninformative. We chose 80 as a support cutoff as 95 has been shown to be overly conservative (Guindon et al. 2010). Gene tree conflict was visualized using the script phypartspiecharts.py (available from https://github.com/mossmatters/MJPythonNotebooks). We conducted more detailed conflict analyses used for edge comparisons discussed below using pxbp as part of the Phyx package (Brown et al. 2017).

Calculating two-topology gene-wise log-likelihoods

The alternate topologies (supermatrix and coalescent) and data matrices for the vertebrate and carnivory datasets were obtained from the original studies, Chiari et al. (2012) and Walker et al. (2017), respectively. We calculated site-wise log-likelihood scores for the two topologies in RAxML using the GTR+ Γ model of evolution, with the data partitioned by gene. The differences in site-wise log-likelihoods between the candidate topologies were then calculated using scripts available from https://github.com/jfwalker/MGWE.

Maximum gene-wise edge calculations

In addition to pairwise topological comparisons, we also examined the maximum gene-wise edges (MGWE) (Fig. 5-1). For a single gene and a single focal edge, the MGWE is the resolution among a set of alternative resolutions for the focal edge that has the highest likelihood from among a set of topologies (more details can be found below). The set of topologies can be determined *a priori* or based on constrained phylogenetic analyses. With this approach, genes are not required to share the same topology even if genes have the same MGWE. This contrasts with a standard shared topology comparison where the topology for each gene would be required to be the same (e.g., supermatrix vs. coalescent topology). Therefore, the MGWE approach allows for genes to have conflicting relationships outside of the edge of interest whether or not they agree with the resolution for the edge of interest. Here, we compared the MGWE for sets of alternative and conflicting edges in order to determine if, by relaxing the requirement for each gene to share the topology, we gain insight into the signal for conflicting relationships.

As mentioned above, the set of topologies that may be used to calculate MGWEs could be determined *a priori* or based on constrained phylogenetic reconstruction analyses. Here, we restricted the tree space under consideration by circumscribing a set of empirically supported topologies (TREESET) consisting of the supermatrix-inferred topology, coalescent inferred topology, and individual gene trees that contained all taxa. For each edge set (i.e., a particular edge and the dominant alternative edges) examined, we pooled trees that were concordant for a particular resolution involving the focal taxa in question for the edge set. Here, for simplicity, we call this set of trees that are concordant for a particular relationship a CADRE. Thus, there was a

CADRE for each resolution for an edge of interest. We then calculated the maximum likelihood for each gene on each topology in the TREESET.

We calculated the MGWEs by retaining the likelihood for the topology with the highest likelihood for each CADRE across all the genes. This became the representative likelihood for that CADRE. The CADRE with the highest likelihood for the gene determined which resolution was the MGWE for that gene.

We then compared this more complex model, allowing for each gene to have a different topology and branch lengths, to the model assuming the shared supermatrix and coalescent topologies. To do this, we calculated the AIC and AICc scores for each CADRE as the summed likelihoods are not comparable given the differences in the number of parameters between the respective models (Theobald 2010; Posada and Buckley 2004). The parameters, k, were calculated based on the number of taxa in each gene, n, and the number of genes in the analysis, g. For a single gene, there were $2\times n-3$ branch length parameters and 9 parameters for the GTR + Γ model of evolution (5 substitution parameters, 1 among-site rate heterogeneity parameter, and 3 estimated base frequencies parameters). The topology was not considered a parameter (Felsenstein, 1983; Yang et al. 1995), when calculating the AIC or AICc scores. The AICc score included a correction for the total number of sites in the supermatrix.

We compared the AIC and AICc scores of several alternative models. First, we ran a standard supermatrix ML analyses assuming a single set of branch lengths on one topology and model parameters unlinked across genes with a GTR + Γ model of evolution ($2\times n-3+9\times g$ parameters). We also conducted a supermatrix analysis allowing the branches to be unlinked across genes including 2 x n – 3 + 9 parameters for every partition and the total parameters being the sum of all parameters for each partition. For this analysis, the number of parameters were the same as those calculated for the CADRE analysis.

Here, we focused on addressing conflicting signal between edges of interest and so the increase in the number of parameters (i.e., a full set for each gene) was considered to be acceptable given our emphasis on gene trees comparisons. However, future work could attempt to limit the expansion of the number of parameters for each CADRE by sharing branch length estimates or model parameters across genes. The code for this analysis is available at https://github.com/jfwalker/MGWE.

Testing for paralogy in carnivory dataset

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

Results

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

Our ML analysis of the vertebrate dataset recovered the same supermatrix topology (Fig. 5-2) as found with ML by Chiari et al. (2012) and Bayesian inference by Brown and Thomson (2017). The difference in log-likelihood between the supermatrix and the coalescent-based Maximum Quartet Support Species Tree (hereafter referred to as coalescent) topologies for the vertebrate dataset was 4.01. Ninety-three of 248 gene trees could be rooted on the outgroup *Protopterus* and only five of these had all taxa represented (Supplementary Table 1). We found low support for relationships within gene trees (SH <80) and substantial gene tree conflict (Fig. 5-2). Of the gene trees with high support (SH >80), seven resolved turtles+crocodilians as sister to birds (hereafter referred to as the vertebrate supermatrix topology) and nine resolved crocodilians+birds sister to turtles (hereafter referred to as the vertebrate coalescent topology).

The two-topology gene-wise log-likelihood comparison showed that 105 genes had a higher likelihood score for the vertebrate supermatrix topology while 143 supported the vertebrate coalescent topology (Figs. 5-3A, 5-4A). Two genes (ENSGALG00000008916 and ENSGALG00000011434, referred to here as 8916 and 11434, respectively), appeared as outliers, exhibiting a disproportionate influence on the overall likelihood of the supermatrix (Fig. 5-3A). The outlier genes identified with maximum likelihood analyses matched those previously identified as outliers using Bayes factors (Brown and Thomson 2017). These two genes both supported the vertebrate supermatrix topology with log-likelihood scores of 79.55 and 46.01 greater than the alternative coalescent tree topology, respectively. The difference in log-likelihood between the two topologies of the non-outlier genes ranged from 0.006 to 19.891 with an average of 3.31 for all genes in the analysis. The removal of the vertebrate genes 8916 and 11434, as shown by Brown and Thomson (2017), recovered the coalescent topology, albeit with low bootstrap support (BS = 12; Appendix D).

Previous work on the carnivory dataset demonstrated that the placement of the *Ancistrocladus+Drosophyllum* clade (Fig. 5-2) contained significant conflict and was strongly influenced by species sampling (Walker et al. 2017). The log-likelihood difference between the supermatrix and coalescent topologies was 74.94 in favor of the former. The two-topology log-likelihood comparison between the dominant topologies on the carnivory dataset (Fig. 5-3B) showed that 623 genes supported *Ancistrocladus+Drosophyllum* sister to all other carnivorous plants (hereafter referred to as carnivory supermatrix topology) while 614 genes supported *Ancistrocladus+Drosophyllum* sister to *Nepenthes alata+Nepenthes ampullaria* (hereafter referred to as carnivory coalescent topology; Figs. 5-3A & 5-4D). Two genes (cluster575 and cluster3300) contributed disproportionately to the overall likelihood. Individually these two genes have a difference in log-likelihood scores between the two topologies of 33.06 and 16.63, respectively, and support the carnivory supermatrix topology. When we reanalyzed the supermatrix with cluster575 and cluster3300 removed, the carnivory coalescent topology was recovered, with 100% BS support (Appendix D). The difference between the two topologies in log-likelihood of the non-outlier genes ranged from 0.001 to 12.82 with an average of 2.82 for all genes in the analysis.

Edge-based analysis

We compared MGWE and two topology gene-wise likelihoods involving the contentious bird, crocodilian, and turtle relationships in the vertebrate dataset (Fig. 5-4B). We found seven unique topologies with the necessary species coverage to conduct the analyses: five gene tree topologies from Chiari et al. (2012) and the two dominant species tree topologies. The set of seven trees included three major conflicting edges for the relationship in question: the two resolutions found in the supermatrix and coalescent trees, and birds sister to crocodilian+mammals+turtles. Ninety-one genes supported the vertebrate supermatrix edge, 144 genes supported the vertebrate coalescent edge, and 13 genes supported the third conflicting edge (Fig. 5-4B). When comparing the supermatrix analysis with a single set of branch lengths, to that where branches are unlinked, we found lower AICc values for unlinked branches (Table 1). The

MGWE AICc scores for the summed likelihoods of the supermatrix (three source trees), the coalescent (three source trees), and the third conflicting edge (one source tree) were highest for the coalescent edge and out of all tested models the coalescent edge was inferred to be the best (Table 1).

For the carnivory dataset, we found 168 unique tree topologies to include in the tree set. The 168 tree topologies contained 45 conflicting edges for the relationship in question with 3 dominant edges. The MGWE analyses found 499 genes supported the supermatrix edge, 466 genes supported the coalescent edge, and 272 genes supported 15 additional edges (Figs. 5-2D, 5-3E). When we further compared the MGWE AICc scores for the supermatrix (44 source trees), the coalescent (56 source trees), and for the third edge (24 source trees) we found the coalescent edge to have the best AICc score out of all tested models (Table 1).

Outlier gene examination

For the carnivory dataset, we explored the possibility that the strongly conflicting genes cluster575 and cluster3300 reflected methodological error in the assembly pipeline, as is the case for the genes identified by Brown and Thomson (2017) for the vertebrate dataset. However, both the alignment and inferred phylogram for each gene revealed no obvious problems or potential sources of systematic error (sparse alignment, abnormally long branch lengths, etc.). We also explored whether compositional heterogeneity could explain the strongly conflicting results (i.e., that the relationships were not truly conflicting, but instead incorrectly modeled). However, both RY-coding in RAxML and explicit modeling of multiple equilibrium frequencies (2, 3, or 4 composition regimes) across the tree in p4 v1.0 (Foster 2004) failed to overturn the inferred relationships. We further explored the possibility of misidentified orthology. The inferred homolog tree produced from amino acid data, containing the outlier gene from the nucleotide dataset, had no signs of misidentified orthology or gene duplication and loss (i.e., an ortholog within the homolog amino acid tree). We found that with the slower amino acid data the sequences in the nucleotide cluster575 were inferred as a single monophyletic ortholog within a duplicated homolog (Appendix D). The discrepancies that appeared between the amino acid dataset and the CDS dataset were found to be either different in-paralogs/splice sites maintained during the dataset cleaning procedure or short sequences that were not identified as homologs in the coding DNA sequence (CDS) dataset (Supplementary Table 2 and

Discussion

Biological processes including substitution saturation, hybridization, horizontal gene transfer, and incomplete lineage sorting can contribute to conflicting signal and may explain both conflict and lack of support widely found in phylogenomic datasets (Salichos et al. 2014; Smith et al. 2015; Kobert et al. 2016). To further complicate the challenges facing phylogenomic analyses, high support values, especially from concatenated analyses, can mask significant underlying conflict (Lee and Hugall, 2003; Ryan et al. 2013; Salichos et al. 2014; Smith et al. 2015; Kobert et al. 2016; Pease et al. 2018). We examined two datasets with extensive conflict involving one or several edges for which small changes in analysis approach or dataset composition altered species tree estimates. Both datasets examined here recovered high support for different topologies based on supermatrix or coalescent species tree analyses.

To address the challenges of conflict and support in phylogenomic datasets, several approaches have been outlined in the literature. In addition to identifying gene tree conflict, these approaches have also highlighted outlier genes that dramatically alter supermatrix analyses (Brown and Thomson 2017; Shen et al. 2017). Both datasets contained genes that exhibited outlier behavior with different topologies inferred depending on the inclusion or exclusion of two genes with disproportionate influence on the likelihood (Brown and Thomson 2017; Walker et al. 2017). In the case of the carnivory dataset, the inferred topology changed with the inclusion or exclusion of just 0.0016% of the genes. The outlier genes in a vertebrate dataset were found to be the result of errors in orthology detection (Brown and Thomson 2017). While the genomic resources were not available to fully examine the carnivorous outlier genes (e.g., we do not yet have synteny or information on gene loss), our analyses did not detect any obvious problems with alignment, compositional heterogeneity, or homology. We found one gene, cluster575, to be an ortholog of a gene that experienced a duplication event prior to the divergence of both ingroup and outgroup taxa (Appendix D). While we could not rule out every possible source of error, we also could not identify a source of methodological error, suggesting the possibility that the disproportionate evolutionary information the gene contains to support the conflicting topology is the result of real (albeit unknown) biological processes.

In addition to the discovery of outlier genes, gene tree analyses and topological examinations have been very informative in the exploration of signal for and against conflicting phylogenetic relationships (Castoe et al. 2009; Smith et al. 2011; Shen et al. 2017). While these analyses can be very helpful in dissecting signal, many assume that a single species tree topology that underlies all genes. For several reasons, this may not be an appropriate model (e.g., hybridization, horizontal gene transfer, and other processes). Conflict among gene trees is common and expected from incomplete lineage sorting, hybridization, and other biological processes. For instance, Jarvis et al. (2014) reported that no gene trees from a genomic data set of 48 species of birds matched the inferred species tree. Furthermore, such a result becomes increasingly likely as sampling breadth (both taxa within a clade as well as the age of the clade itself) increases. The results of a shared-topology analysis may be driven by the resolution of a part of the phylogeny other than the area of interest, as shared-topology analyses condition on fully bifurcating trees that necessarily resolve conflict in the entire tree.

To overcome these limitations, we examined edges across a set of empirically supported candidate topologies, as defined by the set of inferred gene trees and the two dominant species tree hypotheses in question. By examining edges, we accommodate for heterogeneity across the rest of the tree, regardless of the process generating that heterogeneity. The vertebrate gene trees contained three alternative edges for the relationship of interest while the carnivory gene trees contained 45 different edges representing 168 different topologies. Both the MGWE analyses and AICc scores of the vertebrate and carnivory datasets suggested a better fit of the coalescent edge than the supermatrix edge (Table 1). Also, in both cases, we found that the AICc score supported the higher parameterized model, as opposed to a single shared topology and branch lengths. While concatenation is commonly performed using a single set of branch lengths, recent work by Neupane et al. (2018) has also suggested that unlinking branches may be preferred. We do not suggest that the highly parameterized model here is the best model in the universe of possible models, only the best of the ones analyzed.

Our results suggest that future studies may benefit from allowing more heterogeneity than is typically involved in a concatenation analysis. This will require careful examination of the complexity involved in large phylogenomic analyses (e.g., missing data; Stamatakis and Alachiotis 2010). The edge based MGWE analyses facilitate rapid and thorough analysis of the support for relationships across each individual gene. By not conditioning on a single topology for all genes, these analyses can better accommodate the existing heterogeneity between genes while still allowing for edge based investigations. The AIC and AICc analyses allow for more explicit comparisons between the disparate models examined here. Future work could expand on these in several ways. For example, the models explored could potentially have significantly reduced parameters by sharing topologies and branch lengths across some compatible gene regions, including potentially scaling branch lengths proportionally (e.g., as is possible with the -spp option in the program iqtree). Nevertheless, the exploratory analyses presented here provide additional evidence that a simple concatenation approach with these large datasets masks important heterogeneity that can be analyzed further to help inform phylogenetic resolution.

The results presented here contribute to a growing body of literature that addresses how phylogenomic analyses should proceed in the presence of highly influential outlier genes, conflicting topologies, and ever expanding datasets (Wickett et al. 2014; Pease et al. 2016; Brown and Thomson 2017; Shen et al. 2017; Yang et al. 2017). For example, some authors have noted, and it is the case here, that supermatrix analyses may be more susceptible to the problem of strong outliers (Shen et al. 2017; Walker et al. 2017). In these studies, the resolutions inferred using a coalescent method were generally favored. When the dominant process generating gene tree conflict is ILS, coalescent methods should perform better. Some coalescent methods that weigh all gene tree equally (e.g., Mirarab and Warnow 2015), may overcome the problem of outlier genes even if incomplete lineage sorting is not the dominant source of conflict simply by eliminating the disproportionate influence of one or two outlying genes. However, with large and broad datasets, it is more likely that processes in addition to ILS have contributed to gene tree conflict and our ability to accurately reconstruct gene trees may be diminished as we move deeper in the tree of life.

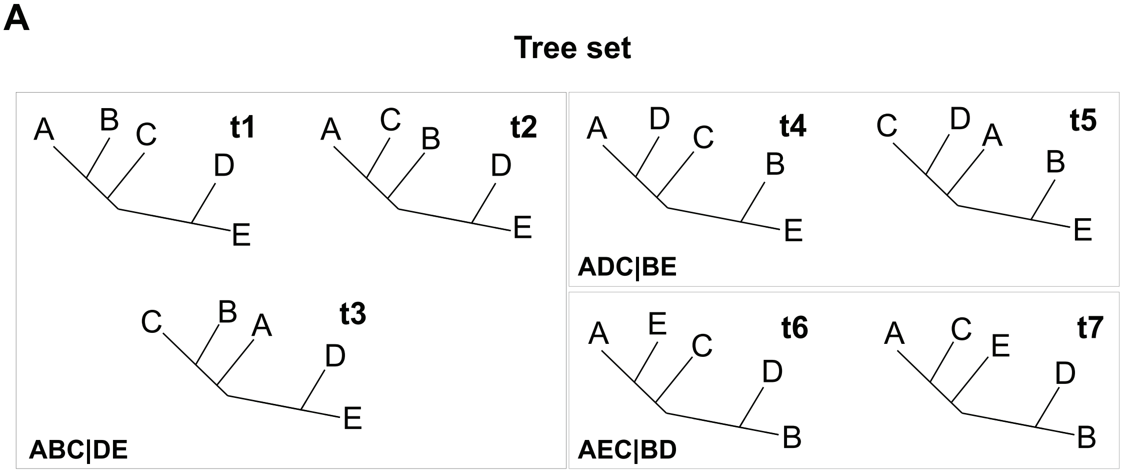
While we continue to uncover the patterns and processes that generate conflicting signal within phylogenomic datasets, it is imperative that we continue to explore ways of dissecting the phylogenetic signal within our datasets. By examining the causes of uncertainty and conflict behind recalcitrant nodes, we can present a more measured confidence, or lack thereof, for particular resolutions. For example, while biological processes most certainly have contributed to the conflict within the datasets examined here, other data set assembly issues (e.g., missing data) may also contribute to conflict and low support in these data sets. For example, while the carnivory dataset had extensive data overlap, the vertebrate dataset only had five gene regions that contained sequence data for every species (Supplementary Table 1). Here we present a framework that focuses on analyzing specific conflicting edges with a MGWE analysis that allows for topological heterogeneity outside of the relationships of interest. This approach accommodates the biological realities of heterogeneity among lineages and throughout a phylogeny in order to address specific questions about an edge of interest. While this is just a small contribution to a growing literature on addressing phylogenomic conflict, as we continue to accommodate more heterogeneity within datasets, we should begin to provide more resolution to important nodes in the tree of life.

Funding

JFW was supported through a fellowship provided by University of Michigan department of Ecology and Evolutionary Biology and the Rackham Pre-doctoral Fellowship. JFW and SAS were supported by NFS 1354048 and JWB and SAS were supported by NSF 1207915.

Acknowledgements

We would like to thank Alexis Stamatakis for advice on RAxML, Caroline Parins-Fukuchi and Nathanael Walker-Hale for critical discussions on the methodology. We would also like to thank Javier Igea, Oscar Vargas, Greg Stull, and Ning Wang for helpful discussion of the manuscript.



B

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
ABC|DE	-20.12 (**t1**)	-10.01 (**t3**)	-11.76 (**t2**)	-14.47 (**t1**)	-9.12 (**t1**)
ADC|BE	-91.42 (**t4**)	-5.15 (**t5**)	-4.19 (**t5**)	-51.25 (**t4**)	-22.47 (**t5**)
AEC|BD	-15.11 (**t6**)	-11.63 (**t7**)	-7.49 (**t6**)	-8.49 (**t7**)	-10.99 (**t6**)

Figure 5-1 Outline for the MGWE procedure

(A) A tree set is depicted with trees numbered. Trees that are concordant for the edge of interest are grouped in boxes with each box representing a CADRE. The concordant edge of interest is denoted at the bottom left-hand corner of each box. (B) A table showing the highest likelihood for each edge calculated from the relevant CADRE and the tree (in parentheses) on which that likelihood was calculated. The MGWE would be the edge for each gene with the highest likelihood.

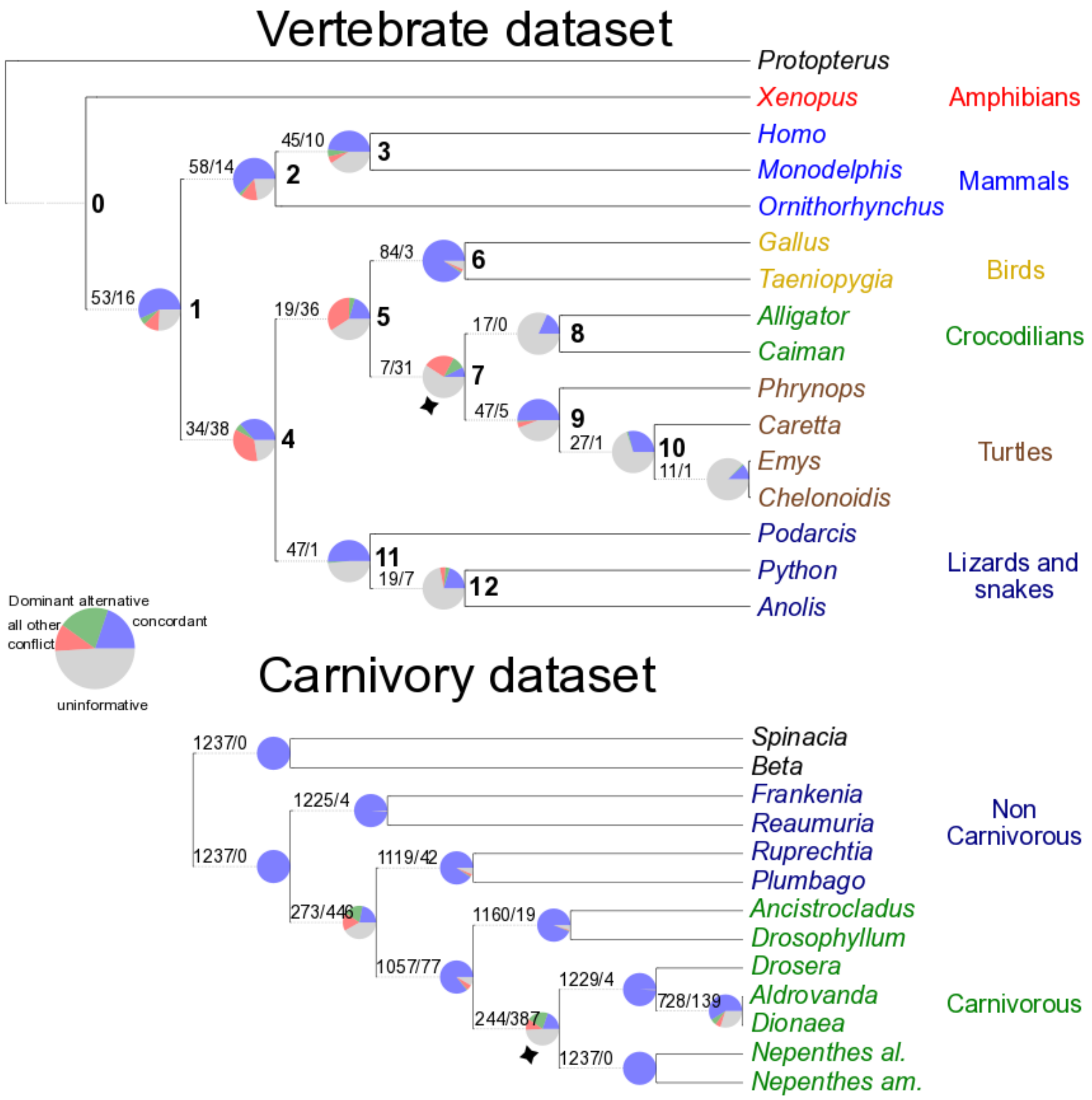


Figure 5-2 Maximum likelihood trees inferred by RAxML for the Chiari et al. 2012 (vertebrate) and Walker et al. 2017 (carnivorous Caryophyllales) datasets.

Conflict analysis for the vertebrate (A) and carnivory (B) datasets. The vertebrate analysis includes the 93 genes that contained the outgroup (*Protopterus*), and the carnivory analysis includes 1237 genes all of which had the outgroups (*Spinacia oleraceae* and *Beta vulgaris*). Black represents gene trees that are concordant with the relationship, the lightest grey represents uninformative genes (SH-like < 80 or no taxon representation for the edge), dark grey represents the dominant alternate topology, and light grey represents all other conflict. Numbers on edges represent concordance/conflict. Bold numbers at the nodes of the vertebrate dataset correspond to edge numbers in Supplementary Table 1.

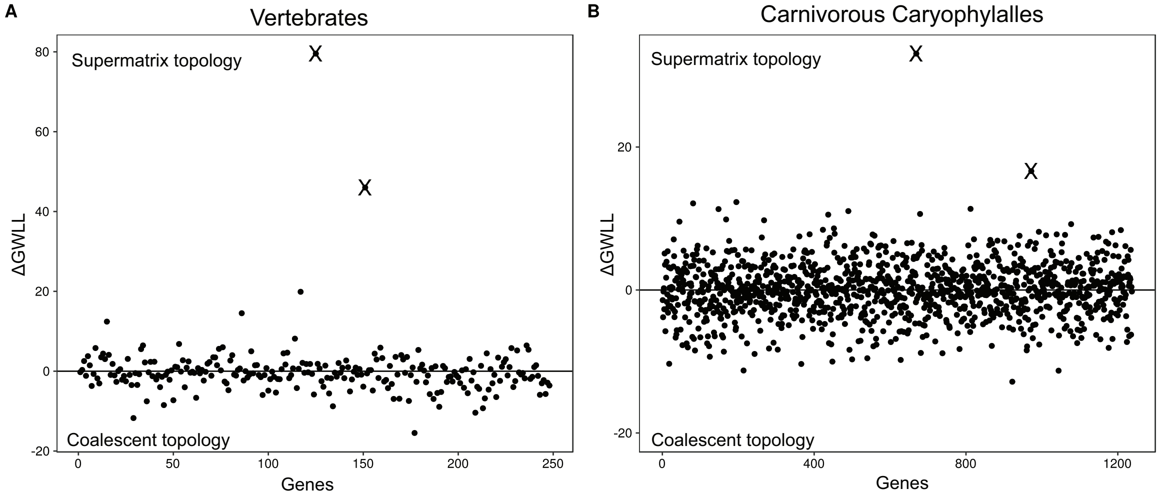


Figure 5-3 Identification of outlier genes using gene-wise likelihood comparison

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

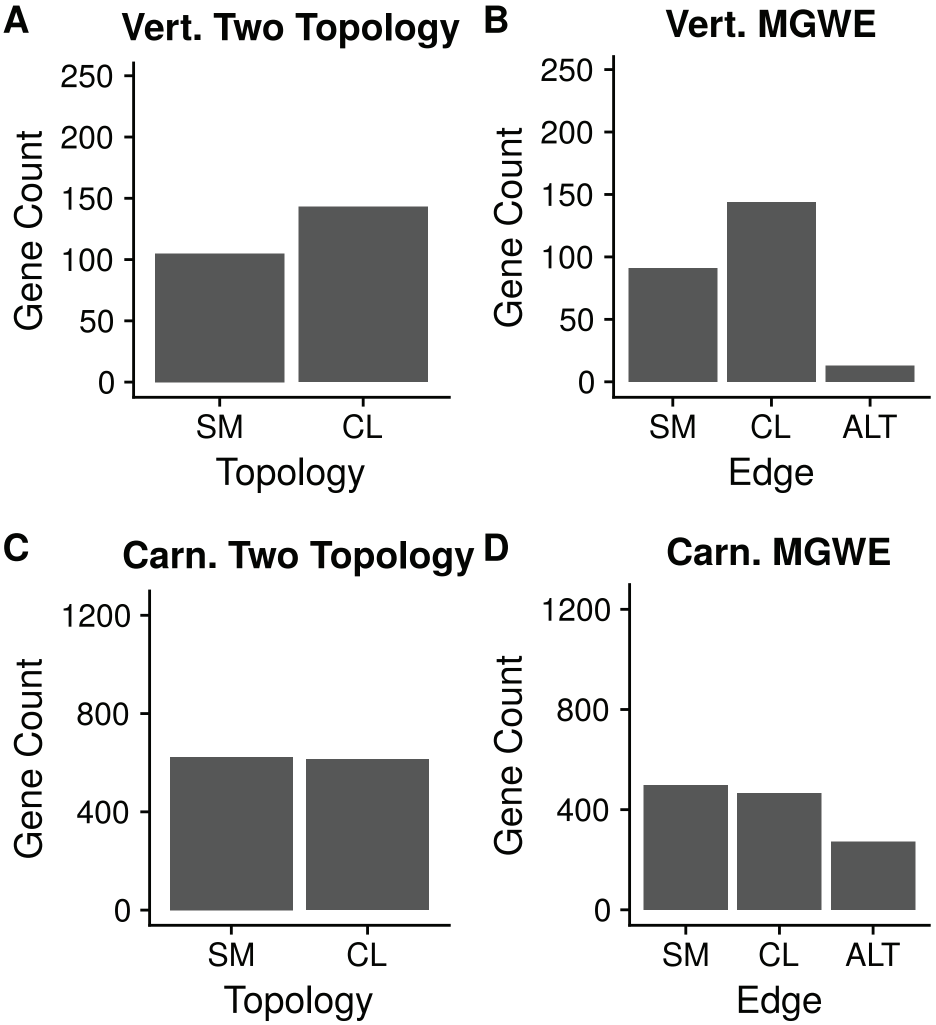


Figure 5-4 Bar plot representing gene counts for the two-topology and MGWE methods

(A and C) The counts of genes that support the supermatrix inferred maximum likelihood (ML) topology and the coalescent-based maximum quartet support species tree (MQSST), for the vertebrate and carnivory datasets respectively. (B and D) The results of the MGWE analysis for support of the edge found in the ML analysis, the conflicting edge from the MQSST analysis, and the sum of all genes supporting an alternative conflict from an edge in the TREESET.

	Relationship	Type	Likelihood	k	AIC	AICc	ΔAICc
Vertebrate	Supermatrix	linked	-1,047,406.05	2261	2,099,334.11	2,099,389.47	21,855.08
		unlinked	-1,031,489.81	7186	2,077,351.63	2,077,925.99	391.59
		Edge	-1,031,423.67	7186	2,077,219.34	2,077,793.70	259.30
	Coalescent	linked	-1,047,410.07	2261	2,099,342.15	2,099,397.51	21,863.11
		unlinked	-1,031,453.35	7186	2,077,278.71	2,077,853.06	318.67
		Edge	**-1,031,294.01**	**7186**	**2,076,960.04**	**2,077,534.39**	**0**
	Dominant Alternative	Edge	-1,041,062.40	7186	2,096,496.81	2,097,071.16	19,536.77
Carnivory	Supermatrix	linked	-13,305,055.20	11156	26,632,422.40	26,632,540.58	35,228.47
		unlinked	-13,261,947.29	39584	26,603,062.59	26,604,570.70	7,258.59
		Edge	-13,258,387.61	39584	26,595,943.24	26,597,451.35	139.24
	Coalescent	linked	-13,305,130.14	11156	26,632,572.28	26,632,690.46	35,378.35
		unlinked	-13,262,019.55	39584	26,603,207.10	26,604,715.22	7,403.10
		Edge	**-13,258,317.99**	**39584**	**26,595,803.99**	**26,597,312.10**	**0**
	Dominant Alternative	Edge	-13,260,106.83	39584	26,599,381.67	26,600,889.78	3,577.67

Table 5-1 Results of model testing the various topologies and edges

In the type column, “linked” represents the supermatrix or coalescent topology with a single set of branch lengths, “unlinked” is the supermatrix or coalescent topology with branch lengths varying among genes, and “Edge” is the MGWE analysis. The top AICc score is bolded.

Chapter VI

Conclusions and future directions

As the field of phylogenomics has matured, it has become more than phylogenetics using genome-scale data. Phylogenomics provides researchers greater insight into the evolutionary history of lineages than simply the reconstruction of relationships. Ancient hybridization, gene duplication and loss, paleopolyploidy, and an array of other biological processes have been inferred using these datasets (Cannon *et al.* 2015; Brockington *et al.* 2015; Yang *et al.* 2015). Comparative functional phylogenomics has now become a standard analysis for any newly sequenced genome and differential gene expression analyses are beginning to be performed in a phylogenetic context (Harkess *et al.* 2017; Dunn *et al.* 2018). Despite these advancements, the methods used to infer species relationships have remained relatively stagnant, with most work focusing on increasing the speed of the same methods used to infer gene trees (Nguyen, Lam-Tung *et al.* 2014).

Increased speed is essential for processing large genomic datasets, especially during the homology identification steps, but a change in underlying methodology should become a major focus. With a greater number of genes included in an analysis, researchers include a more representative view of the complex evolutionary histories within the gene trees—i.e., gene tree conflict (Maddison, 1997; Rokas *et al.* 2003; Smith *et al.* 2015). Given the underlying conflicting signal in any dataset, methods that resolve a single phylogeny are no longer suitable, and increased exploration of conflict is an essential step for the future. My dissertation used the clade

Caryophyllales as model system to advance the field of phylogenomics, focusing in particular on methods analyzing gene tree conflict. A better understanding of the sources of gene tree conflict—and the evolutionary patterns that surround this conflict—will undoubtedly lead to a better understanding of the tree of life.

Caution when using phylogenomics for species tree inference

In evolutionary biology, a phylogeny is typically the starting point for most subsequent analyses (e.g., divergence dating and ancestral state reconstruction). Although it seems the field of phylogenomics is slowly beginning the process of filtering genes for optimal species tree inference. It has been known that many genes do not have sufficient information to properly inform a model of evolution (Yang, 1998), and in phylogenomics this has largely been ignored. With the knowledge that single genes can have such strong influence this is no longer something that can be ignored. If a misalignment induced from incorrectly curated sequence data, can alter inferred species relationships from a 79-gene dataset, then as a field we need to scrutinize the data used. The misalignment resulted in significantly more inferred substitutions in the gene, altering the likelihood score enough to drive the entire dataset toward the wrong topology. Methods that are not influenced by this neglect the biological reality that some genes have greater signal, and therefore the field is in need of novel species tree methods.

Taxon sampling can still alter the estimation of particular relationships with over 1000 genes, and therefore the field needs to start recognizing this again. This was seen in Chapter III, where initial analyses were conducted only on the carnivorous families. However, we were able to get a rare sample from the non-carnivorous family Ancistrocladaceae. When this sample was included in the analyses, it resulted in a significant change in the relationship of the family

Drosophyllaceae with respect to family Nepenthaceae. An investigation of sampling sensitivity revealed that phylogenomic datasets are sensitive to small changes in sampling despite the large dataset sizes. The position of these families directly influences the inference of ancestral states in the group (Heubl *et al.* 2006), and, in particular, how many times these plants are inferred to have evolved from small herbaceous plants into woody lianas.

Overall, it seems the field needs novel methods for species tree inference in phylogenomics. Discovering the relevant biological processes that may contribute to sensitivity to taxon sampling (e.g. hybridization) will be imperative for developing these methods. Perhaps, surprisingly, despite all the sources of conflict, many species relationships are resilient to sampling differences. For phylogenomics, continued on methods specially designed for contentious relationships provide a promising avenue for research.

Methodological error and sources of non-biological gene tree conflict

As discussed throughout this thesis, and in more detail below, there are many biological reasons why a species tree and a gene tree may conflict. To avoid false attribution of biological processes to gene tree conflict, it is essential to determine when the observed conflict has arisen from methodological error. A major source of methodological error is the misidentification of orthology. As phylogenomic datasets grow it will be important to keep using orthology to be determined from phylogenies (Gabaldón, 2008). Misidentified orthology is prevalent in phylogenomics (Brown and Thomson, 2017) and possibly explains some of the highly influential genes across the tree of life (Shen et al. 2017). However, highly influential genes can result from hidden parology (Martin and Burg, 2002), a biological phenomenon. Exploring other ways of determing methodological error versus biological signal is a promising path for the future of phylogenomics. Statistical support can be a way of distinguishing true conflict, however, few species tree methods can factor this in. Finding better ways of incorporating true biologically based conflict will benefit almost all aspects of phylogenomics in the future.

The biological meaning of gene tree conflict in phylogenomic analyses

If a gene tree does not show signs of systematic error, there may be biological processes underlying the conflicting relationships. Although conflict often obscures species relationships, it may also provide valuable information about a lineage's evolutionary history. Conflict can inform researchers of ancient hybridization, gene duplication and loss, horizontal gene transfer, signals of rapid radiations, and a whole array of other past events (Galtier and Daubin, 2008). These phenomena are often perceived as sources of conflict that complicate species tree inference, but they also underscore the reality that speciation may not be the instantaneous, bifurcating process suggested by most phylogenetic trees and that not all gene tree relationships will reflect speciation events (Felsenstein, 1983; Maddison, 1997; Fontaine *et al.* 2015).

With new data sources, researchers can now better document the evolutionary history that is not often reflected in a single species tree. As an analogy, it may be easy for people to distinguish between the colors black and white. However, if shown a gray scale and asked at what point the color changes from white to black, the answer is less clear and varies from person to person. Gene tree conflict shows what is black and what is white, but when combined to form a species tree the definitions mix into grey. When species trees are inferred, it is important to remember the relationships are often not black and white.

A framework for using gene tree conflict to inform species relationships

Using gene tree conflict to help infer species relationships in phylogenomic analyses may prove valuable in the future. This can be done by focusing on a single edge of the phylogeny, thereby allowing the rest of the tree to vary. This type of approach should be of value in the future, but as it currently stands is only in its infancy. Significant methodological development can help overcome the current caveats of the method presented in this thesis. Currently edge based methods rely on a predetermined set of relationships, which limits the power of the analysis. Also, the ability to only explore a single relationship at a time means the rest of the tree relies on other species tree inference methods. Edge based methods will have a bright future for helping understand the tree of life, but significant work needs to be done before they become a standard practice of phylogenomics.

Should phylogenomics return to phylogenetics?

When DNA sequence data was difficult to acquire, researchers often put significant time into examining the data that was available. As an undergraduate, I spent hours looking at Sanger sequencing data for individual genes. This included examining the length of the alignments, gene trees, and searching for any reasons the sequence may not be reliable. In phylogenomics, the idea seems to have emerged that, with enough data, careful scrutiny of individual sequences is no longer necessary. However, this thesis, along with other recent papers, clearly shows that this is not true. Phylogenomics should employ the same careful scrutiny as phylogenetics. Although, phylogenomics has become more than phylogenetics with genome scale data, some of the lessons learned in phylogenetics should be revisited.

Implications for Caryophyllales

The methods developed in this thesis have sought to shed light on the evolution of the plant clade Caryophyllales. The Caryophyllales represent one of the most ecologically and morphologically disparate groups on the planet. This group contains cacti, quinoa, beets, spinach, and the greatest radiation of carnivorous plants on earth (Givnish, 2015). Furthermore, the group has a cosmopolitan distribution, inhabits almost all ecological niches and contains one of two flowering plant species that live on the continent of Antarctica. As shown in my thesis, the group also has notable levels of paleopolyploidy. In the non-core Caryophyllales alone, all families except two have their own unique paleopolyploidy event.

The taxonomy of Caryophyllales is still highly debated (Byng *et al.* 2016). Using broad phylogenomic sampling, Chapter IV sought to address some of the contentious areas in the phylogeny. One of the most contentious regions is the divergence of the family Stegnospermataceae that, prior to Chapter IV, had never been placed with confidence (Hernández-Ledesma *et al.* 2015). The family was found to be sister to the family Macarthuriaceae, which together are part of a number of species poor families forming a grade leading to the main radiation of the core Caryophyllales. The genus *Agdestis* received strong support for its placement sister to the family Sarcobataceae. This, along with morphology disparity from the family Phytolaccaceae, provided evidence that the group should become a separate family. Thus, in Chapter IV we propose recognizing this genus as a separate family, Agdestidaceae.

Another questionable clade is the family Amaranthaceae, which prior to APGIV did not encompass the Chenopodioideae. The Chenopodioideae were previously considered the family Chenopodiaceae, as it is treated in Chapter IV. The decision to dissolve the Chenopodiaceae was due to the placement of the subfamily Polycnemoideae (Kadereit et al. 2003; Masson and Kadereit, 2013). The Polycnemoideae are highly similar to non-Chenopodioideae (the pre-APGIV family Amaranthaceae); however, their phylogenetic position implies they share a more recent common ancestor with Chenopodioideae. To avoid Amaranthaceae being paraphyletic, the family was broadened to encompass the Chenopodiaceae. In Chapter IV, it is argued that this group should be recognized as three families: Amaranthanceae, Chenopodiaceae (reinstated), and Polycnemaceae (a new family); this would resolve the issues of a paraphyletic Amaranthaceae while also allowing the recognition of the highly distinctive Chenopodioideae clade as a separate family. The sampling in Chapter IV only contained two samples of Polycnemoideae, so more detailed sampling will be required before the establishment of formal nomenclatural changes.

Beyond Amaranthaceae, there are many taxonomic problems that will persist despite the use of phylogenomic data. However, instances of paraphyly provide an exciting opportunity to explore how pleisiomorphic traits, such as those that define Polycnemoideae, are reflected within the genome. Phylogenomic datasets now allow us to examine the connection between trait evolution, paraphyly, and gene tree conflict.

In the Caryophyllales, the woody genus *Pereskia* was verified to be paraphyletic, which is interesting, as morphologically the members are extremely similar. Until molecular data was used, it was assumed that *Pereskia*, which has a woody structure, was monophyletic and sister to the succulent family Cactaceae (the cacti) (Edwards *et al.* 2005). By finding that *Pereskia* was non-monophyletic, researchers uncovered strong evidence that the ancestral state of cacti was woody. The genes that conflict and place *Pereskia* as monophyletic will likely yield insight into how woodiness may transition to succulence. There are several other instances of non-monophyly in Carophyllales that deserve attention regarding character evolution. For example, we found that the non-core Caryophyllales (as defined in chapter III) may in fact be non-monophyletic; additionally, the tree from Chapter IV places the carnivorous clade as sister to the core Caryophyllales with weak support.

The non-core Caryophyllales have not always been placed within the Caryophyllales (Rettig et al. 1992), but the monophyly of this group has not previously been questioned. In previous studies, the group has shown up previously with low statistical support for being monophyletic (Cuénoud et al. 2002; Yang et al. 2015). I believe it is important to note that in Chapter III this group is unquestionably recovered as monophyletic, as it has been in many previous studies (Meimberg, 1999; Heubl, 2006; Yang *et al.* 2017). However, the group was rooted on the core Caryophyllales—a standard rooting procedure for the group—and thus the node that would have shown the group as non-monophyletic was forced to be monophyletic. Considering that many genes support its monophyly, while many clearly do not, our phylogenomic results capture important insights that would be lost in typical phylogenetic analyses not considering gene tree conflict.

These results raise the question of what is a clade in the phylogenomic era. The non-core Caryophyllales are supported as monophyletic by several lines of evidence, including morphological and chemical data. However, work that I have conducted has provided more evidence that the group is likely non-monophyletic. This discrepancy is the result of divergence evolutionary histories of many genes, including genes underlying the defining synapomorphies for a group. It is likely that these genes underwent some process or set of processes that resulted in relationships that conflict with the monophyly of the clade. This leads to the question of how does the field proceed for understanding the tree of life in light of non-monophyletic genomes.

Some predictions of the future of phylogenomics

There is little doubt that the fields of biology that examine model organisms and non-model organisms are beginning to merge. Both have made great strides forward over the past century, but until recently there has been little cross talk. However, significant progress is made when methods move from one field to another. For example, the introduction of phylogeny has overturned broad patterns interspecies gene expression (Dunn et al. 2017), helping uncover the power that arises when these two fields are combined. This suggests that targeted studies of a handful of genes for functional analysis should incorporate evolutionary information including phylogeny, something that is beginning to be done with success (Kelly et al. 2017). Non-model organism studies will also greatly benefit from the inclusion of methods typically used for model organisms. For example, most functional studies on non-model organisms base their results on similarity to orthologous genes shared with model organisms. However, this can, at times, be misleading. Nevertheless, researchers will soon be able to create transgenic lines of their own systems.

As the field of transgenics has moved into non-model organisms, the underlying complexity of evolutionary processes has become evident. New questions such as those involving the evolution of pleisiomorphic traits can begin to be examined in more detail. Convergent gene recruitment in polypheletic traits such as C4 photosynthesis, plant carnivory, and succulence can be used as natural replicates for studying trait evolution. This can help guide all of biology to a new era where natural variation amongst species is used to inform hypotheses regarding a species of interest.

Furthermore, extensive genome sequencing of non-model organisms will greatly improve our understanding of genome evolution (Cheng, 2018). We can start resolving the complex connection between micro- and macroevolution. Phylogenomics can help lead us into a new age where differential gene expression, genomic changes, and comparisons among transgenic organisms can uncover evolution through a lens never before imagined. We are entering one of the most exciting times in history to be a systematist.

Appendix A

Supplementary Methods, Figures, and Tables for Chapter II

Supplementary Methods

Statistical Analyses

Because the performance of each gene consists of an aggregate sample of trials (with each node being a trial with outcomes of either concordance or discordance), we analysed the relationships between gene performance and alignment length, tree length, and root-to-tip variance using logistic regression of aggregate binomial trials with the function glm() in R (R Core Team, 2018). Binomial models were generally characterised by high residual deviance, and we thus allowed for overdispersion by fitting quasibinomial logistic regressions (using 'family = quasibinomial()' in R). All code used for these analyses is available on GitHub (https://github.com/jfwalker/ChloroplastPhylogenomics).

We modelled 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 modelled gene performance as a function of length and tree length (as a proxy of gene information content, see **Methods**), to assess the relationship between alignment length and gene performance after controlling for variation associated with gene information content. This has the added benefit of controlling for possible multicollinearity introduced by the covariation between tree length and root-to-tip variation.

Investigation of model fits on full datasets revealed that several observations were highly influential based on leverage and Cook's distance values. Generally, these observations could be predicted based on their outlying values across predictor variables. Therefore, we also conducted investigations on reduced datasets to investigate the influence of these observations. In amino acid datasets, we excluded rpl22 and rpl32, which were probably influential based on their high tree length values, and ycf1 and ycf2, which were probably influential based on their extraordinarily long alignment lengths. In codon datasets, we excluded ndhD, psbL, rpl2 and rpl16, which were probably influential due to high tree length and root-to-tip variance, and ycf1 and ycf2, likewise due to alignment length. Notably, ndhD and rpl2 were also detected as outlier genes (see **Results**). In nucleotide datasets, we excluded clpP, rps15, ycf1 and ycf2.

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 from amino acid and nucleotide datasets, but still excluded ndhD, psbL, rpl2 and rpl16, along with ycf1 and ycf2 from codon datasets.

Regression results were summarised in tables using the R package Stargazer (Hlavac, 2018).

Saturation Analysis

We performed saturation analyses on all the chloroplast genes to determine if they were capable of inferring deep divergence times (Phillipe and Forterre, 1999). 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 et al. 2017), with the JC69 model of evolution used for the correction. The analysis was conducted on the entire gene and on each codon position separately. With the exception of several poorly aligned genes (discussed above), none of the genes analyzed showed significant signatures of saturation (Fig S1).

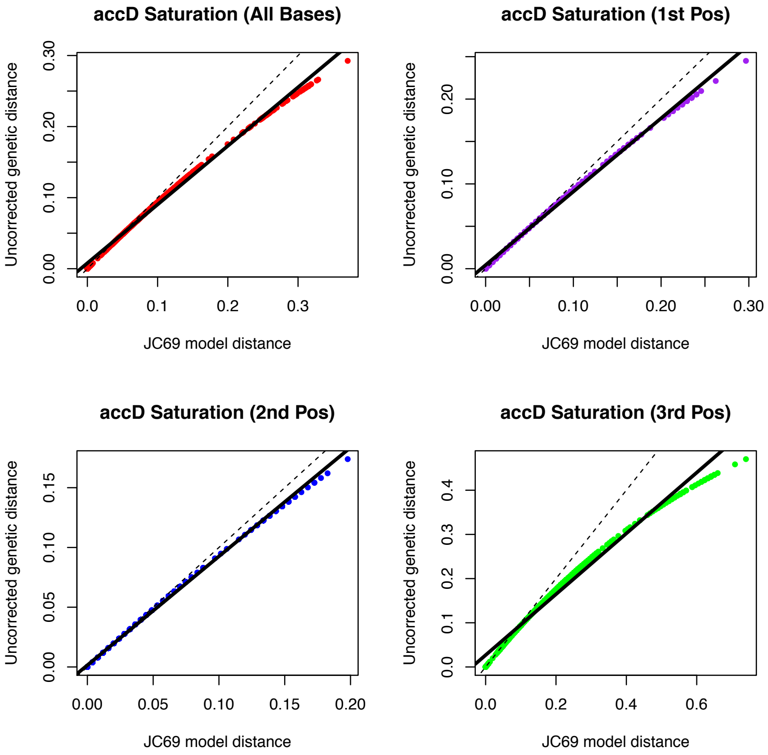


Figure A-1 Saturation plots

Saturation plots for each gene showing the observed number of differences vs. the inferred number of differences, using the JC69 model of evolution for correction.

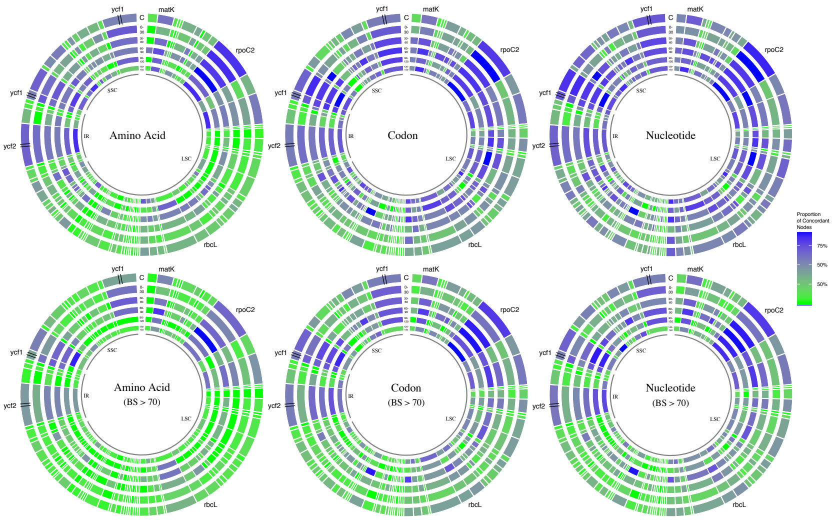


Figure A-2 Concordance mapped based on genomic location

Plastomes diagrams showing concordance levels of each gene (mapped according to their genome positions); gene lengths in the diagrams correspond to alignment lengths. The layers in each diagram represent the gene concordance levels at each of the five time slices. Results are shown for each data type (amino acid, codon, and nucleotide) both with and without bootstrap support.

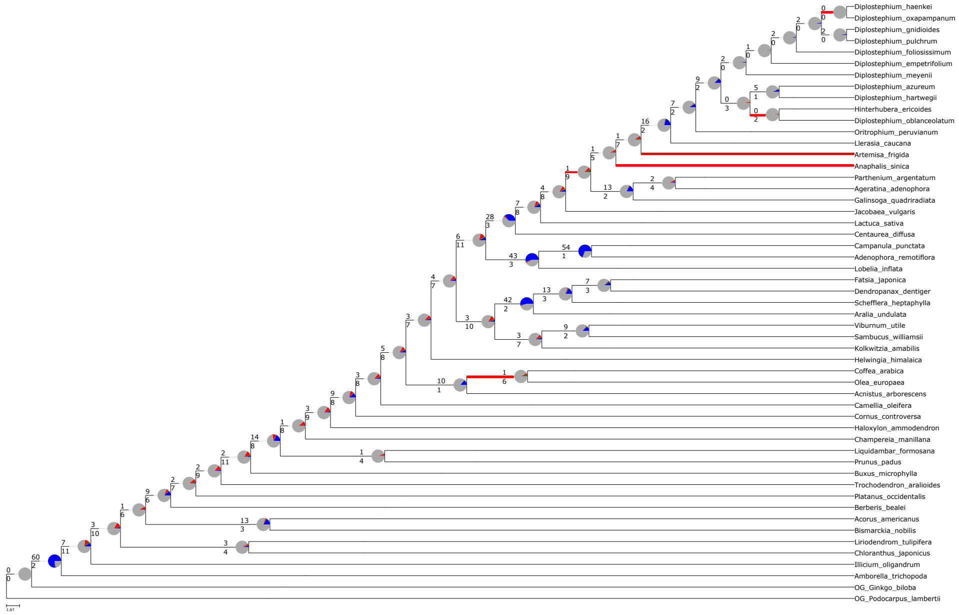


Figure A-3 Conflict of Amino Acids

Conflict analysis of the 'amino acid' gene trees mapped onto the angiosperm reference tree (i.e., the True Topology), including a bootstrap support threshold of 70.

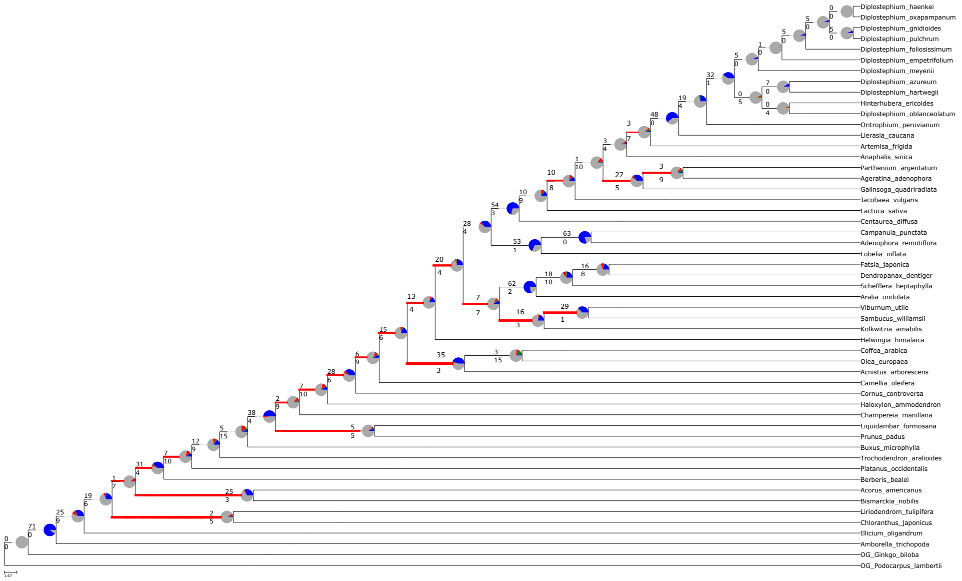
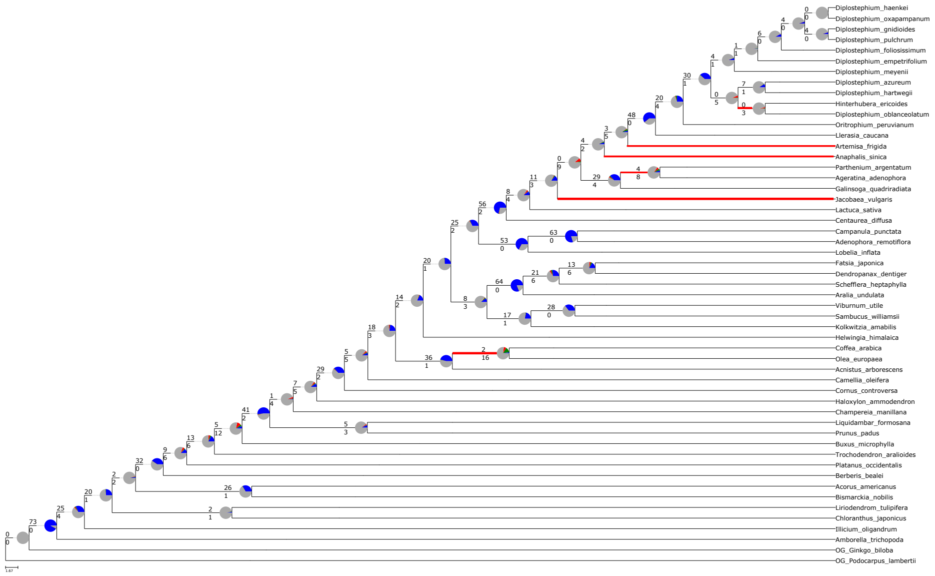


Figure A-4. Conflict of codon aligned gene trees

Conflict analysis of the ‘codon’ gene trees mapped onto the angiosperm reference tree (i.e., the True Topology), including a bootstrap support threshold of 70. reference tree (i.e., the True Topology), including a bootstrap support threshold of 70.



Appendix A-5. Conflict of nucleotide aligned gene trees.

Conflict analysis of the ‘nucleotide’ gene trees mapped onto the angiosperm reference tree (i.e., the True Topology), including a bootstrap support threshold of 70.

Order	Family	Species	Accesion1	Accesion2
Solanales	Solanaceae	Acnistus arborescens	NC_030185.1	KU568472
Acorales	Acoraceae	Acorus americanus	NC_010093.1	EU273602
Asterales	Campanulaceae	Adenophora remotiflora	NC_026999.1	KP889213
Asterales	Asteraceae	Ageratina adenophora	NC_015621.1	JF826503.1
Amborellales	Amborellaceae	Amborella trichopoda	NC_005086.1	AJ506156
Asterales	Asteraceae	Anaphalis sinica	NC_034648.1	KX148081.1
Apiales	Araliaceae	Aralia undulata	NC_022810.1	KC456163
Asterales	Asteraceae	Artemisia frigida	NC_020607.1	JX293720.1
Ranunculales	Berberidaceae	Berberis bealei	NC_022457.1	KF176554
Arecales	Arecaceae	Bismarckia nobilis	NC_020366.1	JX088664
Buxales	Buxaceae	Buxus microphylla	NC_009599.1	EF380351
Ericales	Theaceae	Camellia oleifera	NC_023084.1	JQ975031
Asterales	Campanulaceae	Campanula punctata	NC_033337.1	KU198434
Asterales	Asteraceae	Centaurea diffusa	NC_024286.1	KJ690264.1
Santalales	Opiliaceae	Champereia manillana	NC_034931.1	KY436366
Chloranthales	Chloranthaceae	Chloranthus japonicus	NC_026565.1	KP256024
Gentianales	Rubiaceae	Coffea arabica	NC_008535.1	EF044213
Cornales	Cornaceae	Cornus controversa	NC_030260.1	KU852492
Apiales	Araliaceae	Dendropanax dentiger	NC_026546.1	KP271241

Asterales	Asteraceae	Diplostephium azureum	NC_034882.1	KX063907.1
Asterales	Asteraceae	Diplostephium empetrifolium	NC_034891.1	KX063925
Asterales	Asteraceae	Diplostephium foliosissimum	NC_034883.1	KX063909
Asterales	Asteraceae	Diplostephium gnidioides	NC_034867.1	KX063887
Asterales	Asteraceae	Diplostephium haenkei	NC_034871.1	KX063893
Asterales	Asteraceae	Diplostephium hartwegii	NC_034832.1	KX063880
Asterales	Asteraceae	Diplostephium meyenii	NC_034824.1	KX063919
Asterales	Asteraceae	Diplostephium oblanceolatum	NC_034830.1	KX063941
Asterales	Asteraceae	Diplostephium oxapampanum	NC_034815.1	KX063884
Asterales	Asteraceae	Diplostephium pulchrum	NC_034810.1	KX063857
Apiales	Araliaceae	Fatsia japonica	NC_027685.1	KR021045
Asterales	Asteraceae	Galinsoga quadriradiata	NC_031853.1	KX752097.1
Ginkgoales	Ginkgoaceae	Ginkgo biloba	NC_016986.1	JN867585
Caryophyllales	Amaranthaceae	Haloxylon ammodendron	NC_027668.1	KF534478
Aquifoliales	Helwingiaceae	Helwingia himalaica	NC_031370.1	KX434807
Asterales	Asteraceae	Hinterhubera ericoides	NC_034884.1	KX063910.1
Austrobaileyales	Illiciaceae	Illicium oligandrum	NC_009600.1	EF380354
Asterales	Asteraceae	Jacobaea vulgaris	NC_015543.1	HQ234669.1
Dipsalcales	Caprifoliaceae	Kolkwitzia amabilis	NC_029874.1	KT966716
Asterales	Asteraceae	Lactuca sativa	NC_007578.1	AP007232.1
Saxifragales	Altingiaceae	Liquidambar formosana	NC_023092.1	KC588388
Magnoliales	Magnoliaceae	Liriodendron tulipifera	NC_008326.1	DQ899947

Asterales	Asteraceae	Llerasia caucana	NC_034821.1	KX063908
Asterales	Campanulaceae	Lobelia inflata	NC_033368.1	KY354219.1
Lamiales	Oleaceae	Olea europaea	NC_013707.2	GU228899
Asterales	Asteraceae	Oritrophium peruvianum	NC_034849.1	KX063861
Asterales	Asteraceae	Parthenium argentatum	NC_013553.1	GU120098.1
Proteales	Platanaceae	Platanus occidentalis	NC_008335.1	DQ923116
Pinales	Podocarpaceae	Podocarpus lambertii	NC_023805.1	KJ010812
Rosales	Rosaceae	Prunus padus	NC_026982.1	KP760072
Dipsalcales	Adoxaceae	Sambucus williamsii	NC_033878.1	KX510276
Apiales	Araliaceae	Schefflera heptaphylla	NC_029764.1	KT748629
Throchodendrales	Trochodendraceae	Trochodendron aralioides	NC_021426.1	KC608753
Dipsalcales	Adoxaceae	Viburnum utile	NC_032296.1	KX792264

Table A-1. List of taxa used.

Complete list of taxa included in this study, including corresponding GenBank accession information.

Full Data

	Dependent variable: Total Concordant/Total Discordant					
	logistic					
	AA	AA BS > 70	Codon	Codon BS > 70	Nuc	Nuc BS > 70
Length	0.0002^{***}	0.0002^{***}	0.0004^{***}	0.0003^{***}	0.0002^{***}	0.0002^{***}
	(0.00004)	(0.00004)	(0.00005)	(0.00003)	(0.00002)	(0.00002)
Tree_Length	0.589^{***}	0.701^{***}	-0.009	-0.011	1.306^{***}	1.649^{***}
	(0.046)	(0.063)	(0.009)	(0.011)	(0.104)	(0.127)
Root-to-tip Variance	-41.119^{***}	-74.377^{***}	-0.005^{***}	-0.003	-310.777^{***}	-504.933^{***}
	(5.551)	(10.768)	(0.002)	(0.002)	(34.320)	(43.698)
Constant	-2.012^{***}	-2.968^{***}	-0.458^{***}	-1.304^{***}	-1.618^{***}	-2.732^{***}
	(0.069)	(0.093)	(0.041)	(0.046)	(0.101)	(0.127)
Observations	79	79	79	79	79	79
Log Likelihood	-383.756	-249.108	-422.859	-403.069	-341.953	-312.632
Akaike Inf. Crit.	775.513	506.217	853.719	814.137	691.906	633.263
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table A-2 Logistic regression output using all predictors

Logistic regression output for models including all predictors across all datasets both not considering and considering (BS > 70) support. Parameters are not transformed, i.e., they represent the estimated effect of the predictor on log odds. Quantities in brackets are standard errors.

Full Data

	Dependent variable: Total Concordant/Total Discordant					
	glm: quasibinomial *link = logit*					
	AA	AA BS > 70	Codon	Codon BS > 70	Nuc	Nuc BS > 70
Length	0.0002^{**}	0.0002^{*}	0.0004^{***}	0.0003^{***}	0.0002^{***}	0.0002^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00004)	(0.00004)
Tree_Length	0.589^{***}	0.701^{***}	-0.009	-0.011	1.306^{***}	1.649^{***}
	(0.123)	(0.186)	(0.023)	(0.029)	(0.222)	(0.266)
Root-to-tip Variance	-41.119^{***}	-74.377^{**}	-0.005	-0.003	-310.777^{***}	-504.933^{***}
	(14.787)	(31.532)	(0.004)	(0.006)	(73.500)	(91.484)
Constant	-2.012^{***}	-2.968^{***}	-0.458^{***}	-1.304^{***}	-1.618^{***}	-2.732^{***}
	(0.183)	(0.273)	(0.104)	(0.120)	(0.215)	(0.266)
Dispersion	7.096	8.575	6.569	6.644	4.587	4.383
Observations	79	79	79	79	79	79
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table A-3 Quasibinomial logistic regression output with all predictors

Quasibinomial logistic regression output for models with all predictors across all datasets, both not considering and considering (BS > 70) support. Parameters are not transformed, i.e., they represent the estimated effect of the predictor on log odds. Dispersion gives the estimated quasibinomial dispersion parameter.

Reduced Data

	Dependent variable: Total Concordant/Total Discordant					
	logistic					
	AA	AA BS > 70	Codon	Codon BS > 70	Nuc	Nuc BS > 70
Length	0.002***	0.002***	0.002***	0.002***	0.001***	0.001***
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Tree_Length	0.772***	0.677***	0.732***	0.790***	1.138***	1.138***
	(0.068)	(0.083)	(0.095)	(0.111)	(0.145)	(0.145)
Root-to-tip Variance	-98.944***	-76.416***	-92.568***	-138.623***	-234.793***	-234.793***
	(15.412)	(20.333)	(17.441)	(24.594)	(55.806)	(55.806)
Constant	-2.546***	-3.557***	-1.662***	-2.658***	-1.894***	-1.894***
	(0.083)	(0.111)	(0.106)	(0.127)	(0.113)	(0.113)
Observations	75	75	73	73	75	75
Log Likelihood	-257.248	-156.195	-257.676	-219.627	-253.832	-253.832
Akaike Inf. Crit.	522.496	320.391	523.352	447.253	515.665	515.665
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table A-4 Logistic regression output all predictors excluding influential outliers

Logistic regression results for models with all predictors on datasets excluding influential and outlier observations, both considering and not considering (BS > 70) support. Parameters are not transformed, i.e., they represent the estimated effect of the predictor on log odds. Quantities in brackets are standard errors.

Reduced Data

	Dependent variable: Total Concordant/Total Discordant					
	glm: quasibinomial *link = logit*					
	AA	AA BS > 70	Codon	Codon BS > 70	Nuc	Nuc BS > 70
Length	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}	0.001^{***}	0.001^{***}
	(0.0003)	(0.0002)	(0.0003)	(0.0002)	(0.0001)	(0.0001)
Tree_Length	0.772^{***}	0.677^{***}	0.732^{***}	0.790^{***}	1.138^{***}	1.138^{***}
	(0.129)	(0.094)	(0.166)	(0.173)	(0.242)	(0.242)
Root-to-tip Variance	-98.944^{***}	-76.416^{***}	-92.568^{***}	-138.623^{***}	-234.793^{**}	-234.793^{**}
	(29.067)	(22.918)	(30.688)	(38.290)	(93.245)	(93.245)
Constant	-2.546^{***}	-3.557^{***}	-1.662^{***}	-2.658^{***}	-1.894^{***}	-1.894^{***}
	(0.156)	(0.126)	(0.187)	(0.197)	(0.189)	(0.189)
Dispersion	3.557	1.27	3.096	2.424	2.792	2.792
Observations	75	75	73	73	75	75
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table A-5 Quasibinomial logistic regression all predictors excluding influential outliers

Quasibinomial logistic regression output for all datasets excluding influential observations and outlier genes, both not considering and considering (BS > 70) support. Parameters are not transformed, i.e., they represent the estimate effect of the predictor on log odds. Quantities in brackets are standard errors. Dispersion gives the estimated quasibinomial dispersion parameter.

Full Data

	Dependent Variable: Total Concordant/Total Discordant					
	logistic					
	AA	AA BS > 70	Codon	Codon BS > 70	Nuc	Nuc BS > 70
Length	0.0002***	0.0002***	0.0004***	0.0003***	0.0001***	0.0001***
	(0.00004)	(0.00003)	(0.00005)	(0.00003)	(0.00002)	(0.00001)
Tree_Length	0.266***	0.223***	-0.029***	-0.026***	0.567***	0.412***
	(0.027)	(0.032)	(0.006)	(0.007)	(0.063)	(0.068)
Constant	-1.756***	-2.575***	-0.430***	-1.284***	-1.187***	-1.867***
	(0.060)	(0.076)	(0.039)	(0.045)	(0.087)	(0.097)
Observations	79	79	79	79	79	79
Log Likelihood	-430.988	-295.026	-426.620	-404.311	-390.554	-395.112
Akaike Inf. Crit.	867.975	596.053	859.240	814.622	787.109	796.225
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table A-6 Logistic regression for alignment length and tree length

Logistic regression output for models of alignment length and tree length across all datasets both not considering and considering (BS > 70) support. Parameters are not transformed, i.e., they represent the effect of the predictor on log odds. Quantities in brackets are standard errors.

Full Data

	Dependent Variable: Total Concordant/Total Discordant					
	glm: quasibinomial					
	link = logit					
	AA	AA BS > 70	Codon	Codon BS > 70	Nuc	Nuc BS > 70
Length	0.0002**	0.0002**	0.0004***	0.0003***	0.0001***	0.0001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00004)
Tree_Length	0.266***	0.223***	-0.029**	-0.026	0.567***	0.412**
	(0.076)	(0.072)	(0.015)	(0.019)	(0.166)	(0.176)
Constant	-1.756***	-2.575***	-0.430***	-1.284***	-1.187***	-1.867***
	(0.168)	(0.173)	(0.101)	(0.115)	(0.229)	(0.249)
Dispersion	7.827	5.175	6.575	6.562	6.862	6.621
Observations	79	79	79	79	79	79
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table A-7 Quasibinomial logistic regression for alignment length and tree length

Quasibinomial logistic regression output for models of alignment length and tree length across all datasets both not considering and considering (BS > 70) support. Parameters are not transformed, i.e., they represent the effect of the predictor on log odds. Quantities in brackets are standard errors. Dispersion gives the estimated quasibinomial dispersion parameter.

Reduced Data

	Dependent Variable: Total Concordant/Total Discordant					
	logistic					
	AA	AA BS > 70	Codon	Codon BS > 70	Nuc	Nuc BS > 70
Length	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}	0.001^{***}	0.001^{***}
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.00005)	(0.00005)
Tree_Length	0.291^{***}	0.277^{***}	0.343^{***}	0.260^{***}	0.509^{***}	0.414^{***}
	(0.029)	(0.035)	(0.061)	(0.068)	(0.067)	(0.075)
Constant	-2.390^{***}	-3.414^{***}	-1.390^{***}	-2.292^{***}	-1.580^{***}	-2.516^{***}
	(0.077)	(0.105)	(0.092)	(0.108)	(0.095)	(0.113)
Observations	77	77	73	73	77	77
Log Likelihood	-296.628	-177.738	-272.788	-239.386	-279.551	-248.075
Akaike Inf. Crit.	599.256	361.476	551.577	484.773	565.103	502.151
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table A-8 Logistic regression for alignment length and tree length excluding outliers

Logistic regression output for models of alignment length and tree length on reduced datasets excluding outlier genes and influential observations. Parameters are not transformed, i.e., they represent the effect of the predictor on log odds. Quantities in brackets are standard errors.

Reduced Data

	Dependent Variable: Total Concordant/Total Discordant					
	glm: quasibinomial *link = logit*					
	AA	AA BS > 70	Codon	Codon BS > 70	Nuc	Nuc BS > 70
Length	0.002***	0.002***	0.002***	0.002***	0.001***	0.001***
	(0.0003)	(0.0002)	(0.0003)	(0.0002)	(0.0001)	(0.0001)
Tree_Length	0.291***	0.277***	0.343***	0.260**	0.509***	0.414***
	(0.059)	(0.046)	(0.113)	(0.116)	(0.121)	(0.127)
Constant	-2.390***	-3.414***	-1.390***	-2.292***	-1.580***	-2.516***
	(0.158)	(0.137)	(0.171)	(0.184)	(0.171)	(0.190)
Dispersion	4.25	1.684	3.47	2.918	3.277	2.841
Observations	77	77	73	73	77	77
Note:	*p<0.1; **p<0.05; ***p<0.01					

Table A-9 Quasibinomial logistic regression for alignment length and tree length excluding outliers

Quasibinomial logistic regression output for models of alignment length and tree length across all datasets both not considering and considering (BS > 70) support. Parameters are not transformed, i.e., they represent the estimated effect of the predictor on the log odds. Quantities in brackets are standard errors. Dispersion gives the estimated quasibinomial dispersion parameter.

Appendix B

Supplementary Figures and Tables for Chapter III

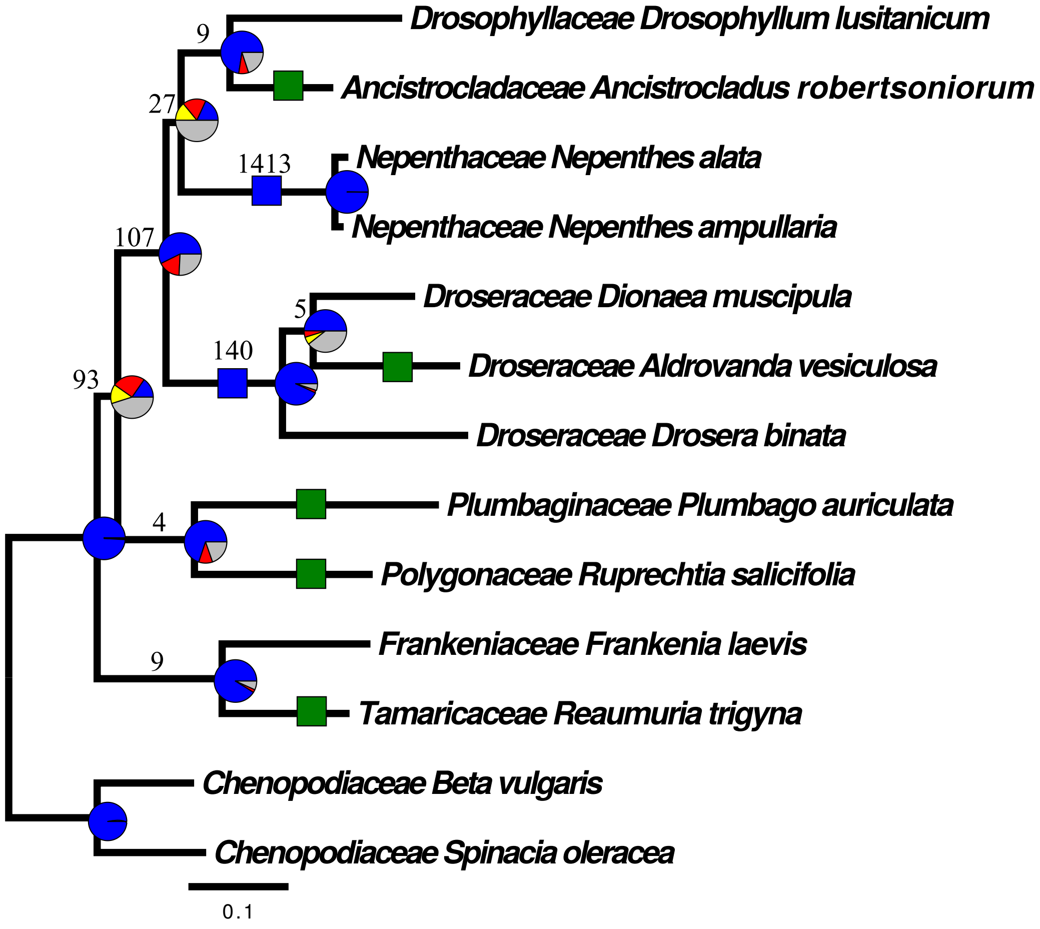

Drosophyllaceae Drosophyllum lusitanicum
Ancistrocladaceae Ancistrocladus robertsoniorum
Nepenthaceae Nepenthes alata
Nepenthaceae Nepenthes ampullaria
Droseraceae Dionaea muscipula
Droseraceae Aldrovanda vesiculosa
Droseraceae Drosera binata
Plumbaginaceae Plumbago auriculata
Polygonaceae Ruprechtia salicifolia
Frankeniaceae Frankenia laevis
Tamaricaceae Reaumuria trigyna
Chenopodiaceae Beta vulgaris
Chenopodiaceae Spinacia oleracea
9
27
1413
107
5
140
93
4
9
0.1

Figure B-1 Species tree from RAxML analysis of the ALLTAX AA supermatrix

Numbers on each branch represent inferred shared unique to clade gene duplications. Squares along branches represent inferred genome duplications, position supported only by Ks plots (Green) and position supported by Ks plots along with shared gene duplications (Blue). Pie charts show gene tree conflict evaluations at each node, proportion concordant (Blue), proportion conflicting (Red), dominant alternative topology (Yellow) and unsupported with SH-Like less than 80 (Grey). Ancestral states on branches taken from *Heubl et. al 2006*.

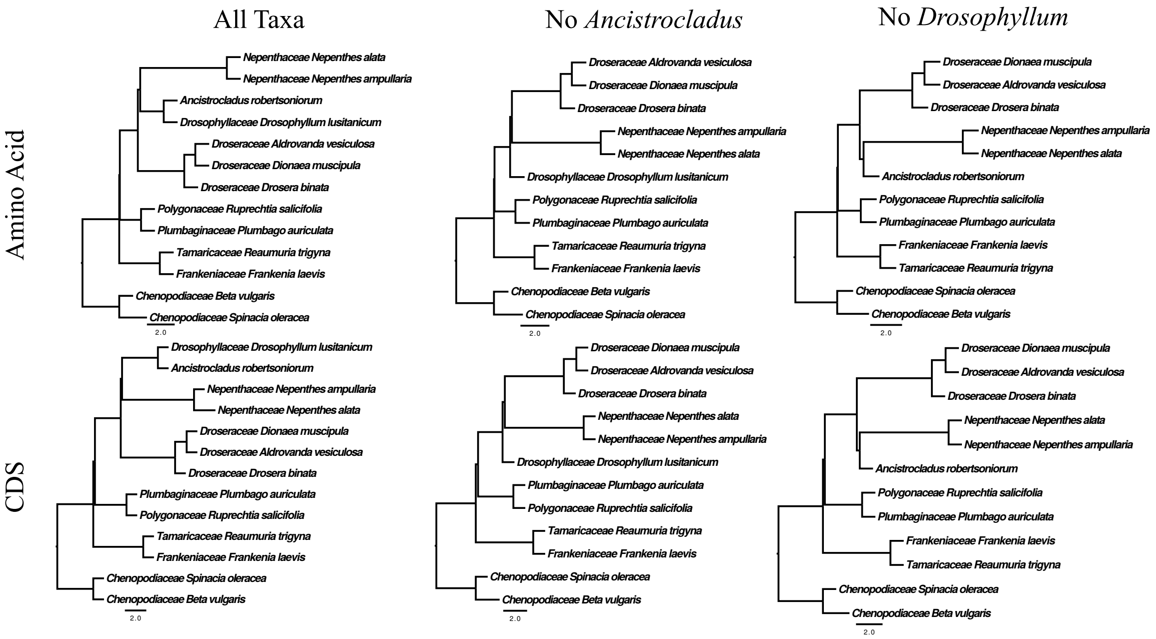
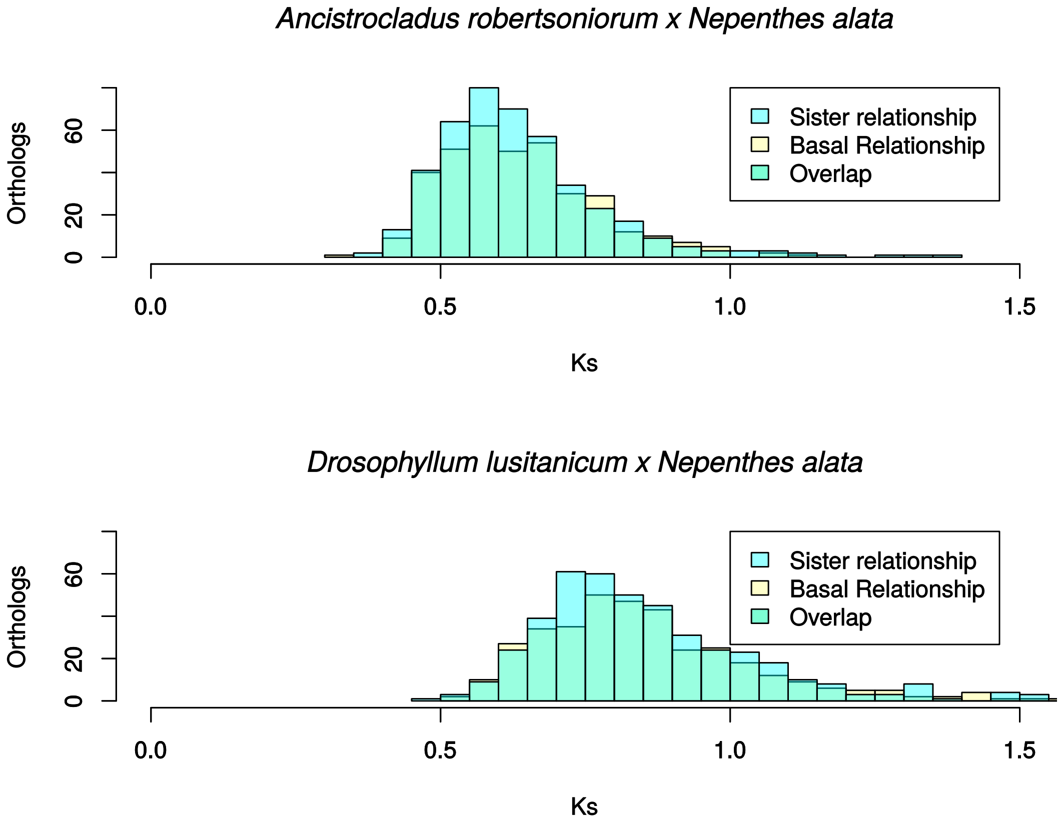

All Taxa
No *Ancistrocladus*
No *Drosophyllum*
Amino Acid
CDS
Nepenthaceae Nepenthes alata
Nepenthaceae Nepenthes ampullaria
Ancistrocladus robertsoniorum
Drosophyllaceae Drosophyllum lusitanicum
Droseraceae Aldrovanda vesiculosa
Droseraceae Dionaea muscipula
Droseraceae Drosera binata
Polygonaceae Ruprechtia salicifolia
Plumbaginaceae Plumbago auriculata
Tamaricaceae Reaumuria trigyna
Frankeniaceae Frankenia laevis
Chenopodiaceae Beta vulgaris
Chenopodiaceae Spinacia oleracea
2.0
Droseraceae Aldrovanda vesiculosa
Droseraceae Dionaea muscipula
Droseraceae Drosera binata
Nepenthaceae Nepenthes ampullaria
Nepenthaceae Nepenthes alata
Drosophyllaceae Drosophyllum lusitanicum
Polygonaceae Ruprechtia salicifolia
Plumbaginaceae Plumbago auriculata
Tamaricaceae Reaumuria trigyna
Frankeniaceae Frankenia laevis
Chenopodiaceae Beta vulgaris
Chenopodiaceae Spinacia oleracea
2.0
Droseraceae Dionaea muscipula
Droseraceae Aldrovanda vesiculosa
Droseraceae Drosera binata
Nepenthaceae Nepenthes ampullaria
Nepenthaceae Nepenthes alata
Ancistrocladus robertsoniorum
Polygonaceae Ruprechtia salicifolia
Plumbaginaceae Plumbago auriculata
Frankeniaceae Frankenia laevis
Tamaricaceae Reaumuria trigyna
Chenopodiaceae Spinacia oleracea
Chenopodiaceae Beta vulgaris
2.0
Drosophyllaceae Drosophyllum lusitanicum
Ancistrocladus robertsoniorum
Nepenthaceae Nepenthes ampullaria
Nepenthaceae Nepenthes alata
Droseraceae Dionaea muscipula
Droseraceae Aldrovanda vesiculosa
Droseraceae Drosera binata
Plumbaginaceae Plumbago auriculata
Polygonaceae Ruprechtia salicifolia
Tamaricaceae Reaumuria trigyna
Frankeniaceae Frankenia laevis
Chenopodiaceae Spinacia oleracea
Chenopodiaceae Beta vulgaris
2.0
Droseraceae Dionaea muscipula
Droseraceae Aldrovanda vesiculosa
Droseraceae Drosera binata
Nepenthaceae Nepenthes alata
Nepenthaceae Nepenthes ampullaria
Drosophyllaceae Drosophyllum lusitanicum
Plumbaginaceae Plumbago auriculata
Polygonaceae Ruprechtia salicifolia
Tamaricaceae Reaumuria trigyna
Frankeniaceae Frankenia laevis
Chenopodiaceae Spinacia oleracea
Chenopodiaceae Beta vulgaris
2.0
Droseraceae Dionaea muscipula
Droseraceae Aldrovanda vesiculosa
Droseraceae Drosera binata
Nepenthaceae Nepenthes alata
Nepenthaceae Nepenthes ampullaria
Ancistrocladus robertsoniorum
Polygonaceae Ruprechtia salicifolia
Plumbaginaceae Plumbago auriculata
Frankeniaceae Frankenia laevis
Tamaricaceae Reaumuria trigyna
Chenopodiaceae Spinacia oleracea
Chenopodiaceae Beta vulgaris
2.0

Figure B-2 Inferred species trees from the Maximum Quartet Supported Species Tree analyses as implemented in Astral

The figure shows the different topologies that result from different combinations of molecules and species sampling inferred using the Maximum Quartet Supported Species Tree (MQSST) as implemented in Astral.



Ancistrocladus robertsoniorum x Nepenthes alata

Drosophyllum lusitanicum x Nepenthes alata

Figure B-3 Distribution of synonymous substitutions (Ks values) among conflicting gene tree topologies

Figure shows the distribution of synonymous substitutions between *Nepenthes alata* and *Ancistrocladus robertsoniorum* and the distribution of synonymous substitutions between *Drosophyllum lusitanicum* and *Nepenthes alata*. The values were acquired for the *A. robertsoniorum*, *D. lusitanicum* and *N. alata* sequences obtained from gene trees that show conflicting topologies of *Drosophyllum* and *Ancistrocladus* sister to *Nepenthes* and *Drosophyllum* and *Ancistrocladus* basal to the rest of the carnivorous Caryophyllales. The mean Ks values for the comparison of *A. robertsoniorum* and *N. alata* were 0.63592 (sister to the other lineages) and 0.6358 (sister to only Nepenthaceae). The mean Ks values for the comparison of *D. lusitanicum* and *N. alata* were 0.85467 (sister to the other lineages) and 0.85861 (sister to Nepenthaceae only).

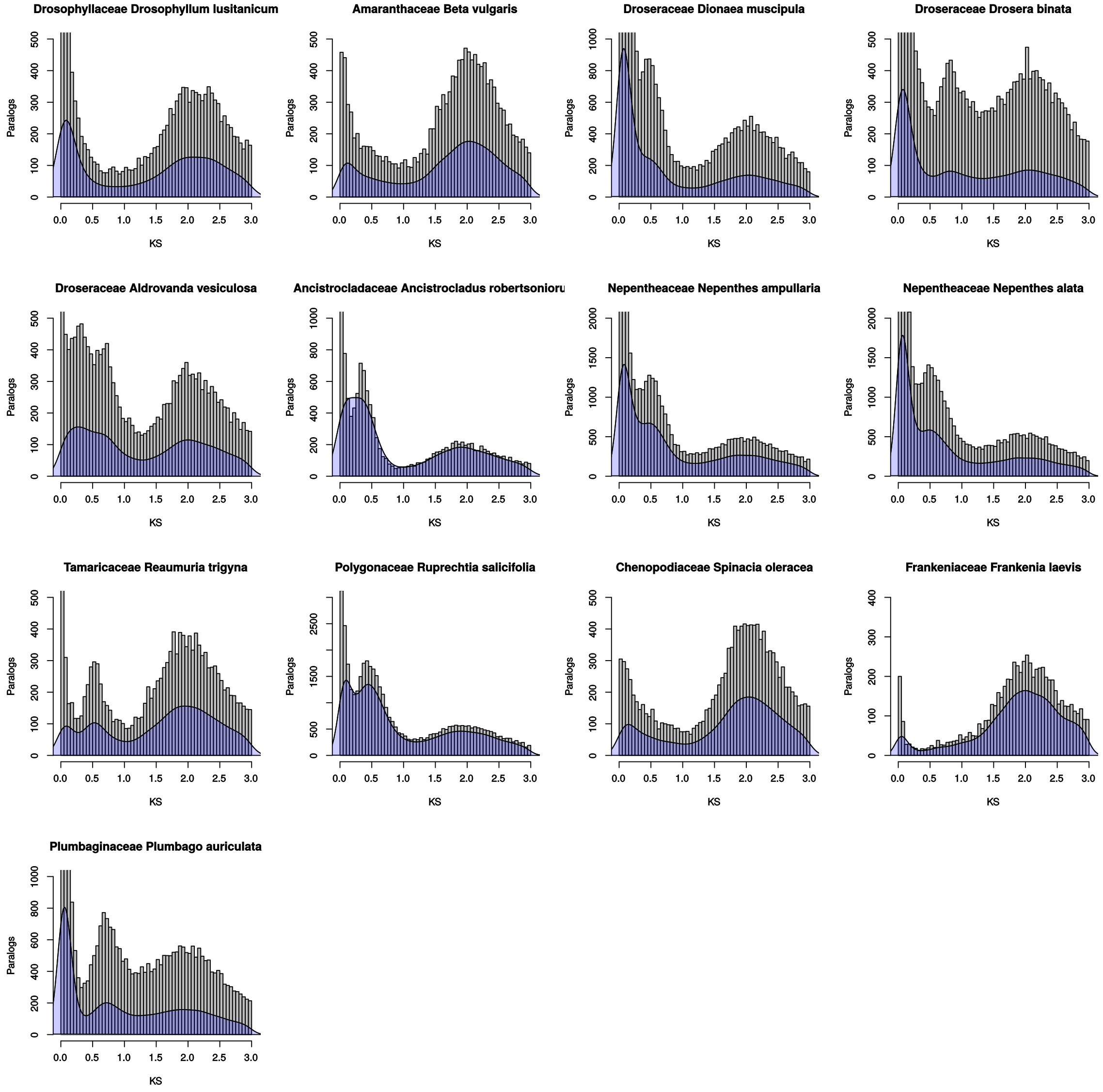


Figure B-4 Comparison of synonymous substitutions (Ks values) between inferred paralogs, presented through a histogram (60 bins) with the density plot mapped on top

Comparison of the within species inferred paralogs Ks values as presented in a histogram of 60 breaks and through a superimposed density plot in blue. The Y-axis is for the histograms representing the paralogs with the given Ks value and the Y-axis for the superimposed density plots is not shown. The X-axis represents the Ks value and is the same between the histogram and the density plot.

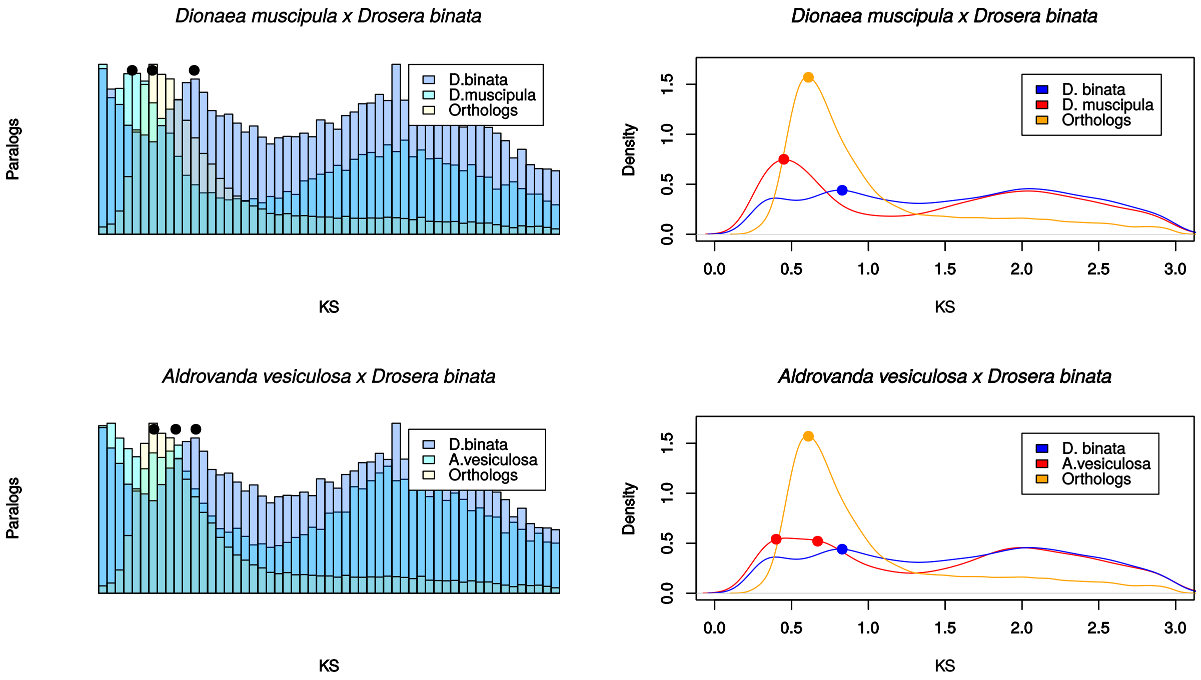


Figure B-5 Synonymous substitution (Ks) plots presented as both histogram and density plot for pairwise Droseraceae comparisons

The figure depicts Ks plots between *Drosera binata* and other members of the Droseraceae. Dots are placed on the highest points of the peaks.

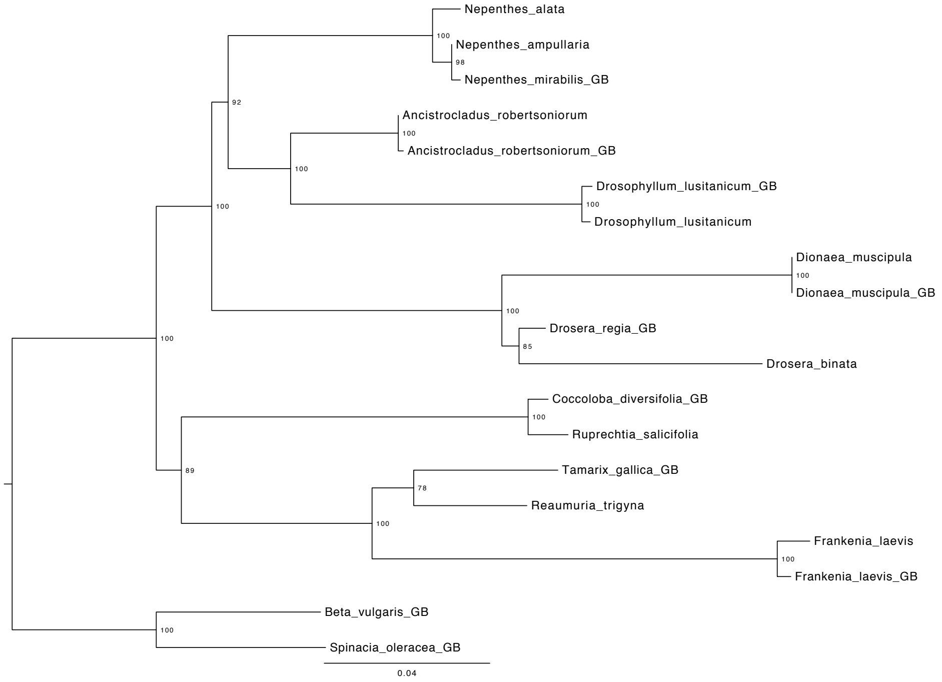


Figure B-6 Contamination check of the transcriptomes through the assembly of a maximum likelihood *matK* gene tree

The figure shows representative family samples from GenBank (ending in GB) compared the MatK sequence inferred using BLAST from the assembled transcriptome data used in the analyses. The analysis was run for 200 BS replicates with the respective values at the nodes.

Taxon Code	**Source (# Reads)**	**# Sequences (gene coverage to *Beta vulgaris*)**	**Collection #**	**Library prep and sequencing**	**Taxon Name and Authority**	**RNA**	**Made for this study**	**Collection Locality**
Beta	http://bvseq.molgen.mpg.de/index.shtml	29,088	N/A	N/A	*Beta vulgaris* (Linnaeus, Carl von)	N/A	No	N/A
Spol	http://bvseq.molgen.mpg.de/index.shtml	23,688	N/A	N/A	*Spinacia oleracea* (Linnaeus, Carl von)	N/A	No	N/A
WPYJ	http://dx.doi.org/10.5061/dryad.33m48	17,678 **(8,218)**	N/A	N/A	*Frankenia laevis* (Linnaeus, Carl von)	N/A	No	N/A
Retr	SRX105466 & SRX099851 **(13,633,333 & 12,777,778)**	26,934 **(9,496)**	N/A	N/A	*Reaumuria trigyna* (Maximowicz, Carl Johann (Ivanovič))	N/A	No	N/A
RuprSFB	BioSample: SAMN05936389 Experiment: SRX2268492 Run: SRR4450414 **(40,463,293)**	65,889 **(9,135)**	N/A	TruSeq Stranded; HiSeq 2000 Paired End 101 bp multiplex 7 per lane	*Ruprechtia salicifolia* (Meyer, Carl Anton (Andreevič) von)	Purelink	Yes	Cultivated at Cambridge University Botanic Garden
MJM3360	BioSample: SAMN05936390 Experiment: SRX2268493 Run: SRR4450415 **(28,674,244)**	63,905 **(8,887)**	Michael J. Moore et al. 3360 (OC)	KAPA stranded; HiSeq 4000 Paired end 150 bp multiplex 11 per lane	*Plumbago auriculata* (Lamarck, Jean Baptiste Antoine Pierre de Monnet de)	Purelink with young leaf, red flower bud	Yes	The Kampong: between tennis court and main house.
DrolusSFB	BioSample: SAMN05936387 Experiment: SRX2268483 Run: SRR4450406 **(37,943,871)**	44,804 **(9,804)**	N/A	KAPA Stranded; HiSeq 2000 Paired End 101 bp multiplex 7 per lane	*Drosophyllum lusitanicum* ((L.) Link)	Purelink with young leaves	Yes	Cultivated at Cambridge University Botanic Garden
MJM2940	BioSample: SAMN05936890 Experiment: SRX2268484 Run: SRR4450407 **(34,779,941)**	58,310 **(10,070)**	Michael J. Moore & J. Lee 2940 (OC)	KAPA stranded; HiSeq2500V4 Paired End 125 bp multiplex 11 per lane	*Ancistrocladus robertsonorium* (J. Leonard)	Purelink with young leaf, apical meristem	Yes	Missouri Botanical Garden, cultivated in Climatro

						(red)		n for years.
NepSFB	BioSample: SAMN05936153, SAMN05936157, SAMN05936158 Experiment: SRX2268491, SRX2268490, SRX2268488 Run: SRR4450413, SRR4450412, SRR4450410 **(27,015,843, 27,848,222, 39,264,059)**	96,679 **(9,523)**	N/A	KAPA stranded; HiSeq 2000 Paired End 101 bp multiplex 7 per lane	*Nepenthes alata* (Blanco, Francisco Manuel)	Purelink with small, median, and large traps respectively	Yes	Cultivated at Cambridge University Botanic Garden.
Neam	SRR2866506, SRR2866512, SRR2866533 **(22,803,439, 21,322,400, 26,615,789)**	84,007 **(9,446)**	N/A	HiSeq 2000 Paired End 101 bp multiplex 7 per lane	*Nepenthes ampullaria* (Jack, William)	N/A	No	N/A
Dino	SRX1376794 **(40,159,392)**	88,684 **(8,621)**	N/A	N/A	*Dionaea muscipula* (Ellis, John)	N/A	No	N/A
MJM1652	SRR1979677 **(25,365,347)**	26,040 **(8,487)**	N/A	N/A	*Aldrovanda vesiculosa* (Linnaeus, Carl von)	N/A	No	N/A
DrobinSFB	BioSample: SAMN05936370, SAMN05936384, SAMN05936385 Experiment: SRX2268489, SRX2268486, SRX2268487 Run: SRR4450411, SRR4450408, SRR4450409 **(36,941,729, 36,126,728, 36,508,405)**	65,080 **(7,605)**	N/A	HiSeq 2000 Paired End 101 bp multiplex 7 per lane	*Drosera binata* (Labillardière, Jacques Julien Houtou de)	Purelink with small, median, and large traps respectively	Yes	Cultivated at Cambridge University Botanic Garden.

Table B-1 Taxa used for the analyses, number of sequences and sequence coverage, sources of data, collections and location

Species	GenBank Accession
Drosera regia	gi|8568032|gb|AF204848.1|
Dionaea muscipula	gi|8568030|gb|AF204847.1|
Nepenthes mirabilis	gi|14193614|gb|AF315920.1|
Tamarix gallica	gi|8568058|gb|AF204861.1|
Frankenia laevis	gi|47498931|gb|AY514853.1|
Coccoloba diversifolia	gi|297372635|emb|FN597640.1|
Ancistrocladus robertsoniorum	gi|285803889|gb|GQ470539.1|
Drosophyllum lusitanicum	gi|47498945|gb|AY514860.1|
Beta vulgaris	gi|47498889|gb|AY514832.1|
Spinacia oleracea	gi|11497503:1783-3300
Drosera regia	gi|8568032|gb|AF204848.1|
Dionaea muscipula	gi|8568030|gb|AF204847.1|
Nepenthes mirabilis	gi|14193614|gb|AF315920.1|

Table B-2 List of species and GenBank accession for the MatK sequences used in the contamination analysis

Species	GenBank Accession	Function
Arabidopsis thaliana	gi|42568444|ref|NP_199851.2|	Purple Acid Phosphotase 27
Arabidopsis thaliana	gi|1032282051|gb|OAO96379.1|	Cysteine peptidase C1A (SAG12)
Arabidopsis thaliana	gi|15230262|ref|NP_191285.1|	Beta-Glucanase (BGL2)
Arabidopsis thaliana	gi|1032291674|gb|OAP06001.1|	Serine Carboxypeptidase 49 (SCPL49)
Arabidopsis thaliana	gi|1032297141|gb|OAP11467.1|	Ribonuclease T2 (RNS1)
Dionaea muscipula	gi|563616779|gb|AHB62682.1|	Chitinase Class I (VF CHITINASE I)
Cucumis sativus	gi|167533|gb|AAA33129.1|	Plant Peroxidase
Arabidopsis thaliana	gi|186500492|ref|NP_001118321.1|	Plant Lipid Transfer Protein
Zea mays	gi|413947720|gb|AFW80369.1|	Peptide-N4-Asparagine Amidase A
Camellia sinensis	gi|558483701|gb|AHA56682.1|	Pathogenesis-related protein
Arabidopsis thaliana	gi|42562696|ref|NP_175606.2|	LysM-containing protein

Cynara cardunculus	gi|976927626|gb|KVI11230.1|	Aspartic peptidase

Table B-3. Samples used for identifying homologous clusters of genes identified to be important in carnivory from *Bemm et. al 2016.* Including species name, GenBank accession and function of sequences.

Gene family name	Size of family	Copies in non-carnivorous taxa	Copies per non-carnivorous taxa	Copies in carnivorous taxa	Copies per carnivorous taxa	Putative function
cluster1_1rr_2rr.fa.mafft.aln	3498	1927	275.28	1571	261.8	putative leucine-rich repeat receptor-like protein kinase At2g19210
cluster3rr_1rr.fa.mafft.aln	3000	1513	216.1	1487	247.8	pentatricopeptide repeat-containing protein At4g02750
cluster2_1rr_1rr.fa.mafft.aln	2479	1350	192.8	1129	188.1	probable LRR receptor-like serine/threonine-protein kinase At2g24230
cluster4_1rrrr.fa.mafft.aln	2479	1321	188.7	1158	193	pentatricopeptide repeat-containing protein At5g15280
cluster6_1rrrr.fa.mafft.aln	1201	658	188.7	543	193	geraniol 8-hydroxylase-like [Citrus sinensis]
cluster7rr_2rr.fa.mafft.aln	1039	536	76.6	503	83.8	CBL-interacting protein kinase 07 [Vitis vinifera]
cluster10rrrr.fa.mafft.aln	762	348	49.7	348	69	29 kDa ribonucleoprotein A, chloroplastic [Eucalyptus grandis]
cluster8rrrr.fa.mafft.aln	757	386	55.1	371	61.8	UDP-glycosyltransferase 84A22 [Camellia sinensis]
cluster12rrrr.fa.mafft.aln	730	348	49.7	382	63.7	probable envelope ADP,ATP carrier protein, chloroplastic [Beta vulgaris subsp. vulgaris]

cluster13rrrr.fa.mafft.aln	638	315	45	323	53.8	probable protein phosphatase 2C 12 [Theobroma cacao]
cluster21_1rr_1rr.fa.mafft.aln	638	307	43.8	331	55.2	transcription factor MYB44-like [Beta vulgaris subsp. vulgaris]
cluster17rr_1rr.fa.mafft.aln	619	310	44.28	309	51.5	probable DEAD-box ATP-dependent RNA helicase 48 isoform X1 [Fragaria vesca subsp. vesca]
cluster22_2rrrr.fa.mafft.aln	619	314	44.8	305	50.8	Ras-related protein RGP1 [Anthurium amnicola]
cluster11rrrr.fa.mafft.aln	602	317	45.3	285	47.5	1-aminocyclopropane-1-carboxylate oxidase homolog 1-like [Vitis vinifera]
cluster18rrrr.fa.mafft.aln	595	341	48.7	254	42.3	GDSL esterase/lipase At1g71691 [Ziziphus jujuba]

Table B-4 List of largest gene families, divided to size of family found in the carnivorous and non-carnivorous taxa used in the study.

Name in analysis	Size of family	Copies in non-carnivorous taxa	Average copies per non-carnivorous taxa	Average copies in carnivorous taxa	Average copies per carnivorous taxa	Putative function
cluster98rrrr.fa.mafft.aln	199	103	14.7	96	16	Purple Acid Phosphotase 27
cluster82rrrr.fa.mafft.aln	234	113	16.1	121	20.1	Cysteine peptidase C1A (SAG12)
cluster32rrrr.fa.mafft.aln	416	214	30.5	202	33.6	Beta-Glucanase (BGL2)
cluster7000rrrr.fa.mafft.aln	8	3	0.4	5	0.8	Serine Carboxypeptidase 49 (SCPL49)
cluster898rrrr.fa.mafft.aln	50	26	3.71	24	4	Ribonuclease T2 (RNS1)
cluster319_2rrrr.fa.mafft.aln	62	37	5.2	25	4.1	Chitinase Class I (VF CHITINASE I)
cluster24rrrr.fa.mafft.aln	527	324	46.2	204	33.83	Plant Peroxidase
cluster263rrrr.fa.mafft.aln	108	47	6.7	61	10.1	Plant Lipid Transfer Protein
cluster1669rrrr.fa.mafft.aln	25	13	1.8	12	2	Peptide-N4-Asparagine Amidase A
cluster556rrrr.fa.mafft.aln	69	44	6.2	25	4.1	Pathogenesis-related protein
cluster6240rrrr.fa.mafft.aln	9	7	1	2	0.3	LysM-containing protein
cluster439rrrr.fa.mafft.aln	70	28	4	42	7	Aspartic peptidase

Table B-5 Comparison of gene family size between carnivorous and non-carnivorous taxa identified in carnivory from *Bemm et. al 2016*

	ALLTAX	NODROS	NOANC
amino acid (AA) homologs	10531	10152	9999
coding DNA sequence homologs	10766	9910	9388
amino acid (AA) orthologs	1637	1616	1614
coding DNA sequence orthologs	1237	1211	1117

Table B-6 Composition of datasets used for the phylogenomic analyses

Appendix C

Supplementary Tables for Chapter IV

Number of sequences: 305		
Sequence length: 15467		
--------Prot TABLE---------		
Prot	Total	Proportion
A	226039	0.0479157
C	43804	0.00928555
D	175148	0.0371278
E	203845	0.043211
F	137153	0.0290736
G	195552	0.041453
H	67011	0.014205
I	163969	0.0347581
K	205600	0.043583
L	314684	0.0667066
M	73269	0.0155315
N	121327	0.0257188
P	157914	0.0334745
Q	119242	0.0252769
R	166134	0.035217
S	270698	0.0573825
T	156366	0.0331464
V	211596	0.044854
W	36205	0.00767472
Y	88029	0.0186604
-	1.58385e+06	0.335743
X	3	6.35939e-07
Prot	Total	Proportion

Table C-1 Statistics and information for the supermatrix of all taxa

Number of sequences: 45		
Sequence length: 348593		
--------Prot TABLE---------		
Prot	Total	Proportion
A	909555	0.0579826
C	226072	0.0144117
D	716925	0.0457028
E	835508	0.0532622
F	544080	0.0346842
G	820544	0.0523083
H	304617	0.0194188
I	680274	0.0433663
K	834522	0.0531994
L	1.27578e+06	0.0813287
M	292242	0.0186299
N	535133	0.0341138
P	677477	0.043188
Q	465094	0.029649
R	699701	0.0446048
S	1.15795e+06	0.0738175
T	669625	0.0426875
V	885747	0.0564649
W	159355	0.0101586
Y	366122	0.0233397
-	2.63029e+06	0.167676
X	75	4.78112e-06
Prot	Total	Proportion

Table C-2. Statistics and information for the family Caryophyllaceae.

Number of sequences: 41		
Sequence length: 177349		
--------Prot TABLE---------		
Prot	Total	Proportion
A	417290	0.0573886
C	97948	0.0134705
D	358235	0.0492669
E	430693	0.0592318
F	245799	0.033804
G	391235	0.0538053
H	154344	0.0212264
I	319853	0.0439884
K	407886	0.0560953
L	614805	0.0845522
M	138586	0.0190593
N	256826	0.0353205
P	314959	0.0433153
Q	237145	0.0326138
R	337049	0.0463533
S	568473	0.0781803
T	302783	0.0416408
V	413076	0.056809
W	76214	0.0104815
Y	165818	0.0228044
-	1.02229e+06	0.140593
X	0	0
Prot	Total	Proportion

Table C-3 Statistics and information for the family Nyctaginaceae

Number of sequences: 47		
Sequence length: 610051		
--------Prot TABLE---------		
Prot	Total	Proportion
A	1.73512e+06	0.0605153
C	384488	0.0134097
D	1.30337e+06	0.0454573
E	1.63689e+06	0.0570893
F	887148	0.0309408
G	1.54966e+06	0.054047
H	534107	0.0186279
I	1.07966e+06	0.0376551
K	1.50406e+06	0.0524567
L	2.13606e+06	0.0744988
M	530616	0.0185062
N	923961	0.0322248
P	1.29361e+06	0.0451171
Q	865486	0.0301853
R	1.35605e+06	0.0472945
S	2.16828e+06	0.0756226
T	1.07176e+06	0.0373796
V	1.50887e+06	0.0526244
W	251897	0.00878535
Y	569724	0.0198701
-	5.38159e+06	0.187692
X	0	0
Prot	Total	Proportion

Table C-4 Statistics and information for the family Cactaceae

Number of sequences: 65		
Sequence length: 149426		
--------Prot TABLE---------		
Prot	Total	Proportion
A	580786	0.0597966
C	129193	0.0133015
D	441805	0.0454874
E	553995	0.0570383
F	350780	0.0361156
G	506476	0.0521458
H	175809	0.018101
I	428804	0.0441488
K	548528	0.0564754
L	800732	0.0824418
M	181817	0.0187195
N	341956	0.0352071
P	427572	0.044022
Q	308713	0.0317845
R	436410	0.0449319
S	727731	0.0749258
T	403996	0.0415947
V	534580	0.0550393
W	104371	0.0107458
Y	222277	0.0228852
-	1.50635e+06	0.155091
X	6	6.17749e-07
Prot	Total	Proportion

Table C-5 Statistics and information for the families Amaranthaceae and Chenopodiaceae

Number of sequences: 60		
Sequence length: 209288		
--------Prot TABLE---------		
Prot	Total	Proportion
A	750145	0.0597379
C	165423	0.0131735
D	570064	0.0453971
E	690076	0.0549543
F	410840	0.0327173
G	659010	0.0524803
H	225925	0.0179916
I	500390	0.0398486
K	636219	0.0506654
L	1.00068e+06	0.0796896
M	224102	0.0178464
N	379778	0.0302437
P	524714	0.0417856
Q	380877	0.0303312
R	583565	0.0464722
S	940418	0.0748903
T	468891	0.0373402
V	665667	0.0530104
W	123957	0.00987133
Y	269341	0.021449
-	2.38719e+06	0.190104
X	0	0
Prot	Total	Proportion

Table C-6 Statistics and information for the family Nyctaginaceae

Species	Credit	Figure
Nitrophila occidentalis	Michael J. Moore	5
Beta vulgaris	By Evan-Amos - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=16069395	5
Spinacia oleraceae	By Victor M. Vicente Selvas - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=11971683	5
Amaranthus tricolor	by Kurt Stueber, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=7411	5
Grayia spinosa	By Stan Shebs, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1677841	5
Ferocactus latispinus	Lucas C. Majure	6
Opuntia arenaria	Lucas C. Majure	6
Pereskia grandiflora	By Kurt Stüber [1] - caliban.mpiz-koeln.mpg.de/mavica/index.html part of www.biolib.de, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6891	6
Colobanthus quitensis	By Liam Quinn - Flickr: Antarctic Pearlwort, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=15525940	7
Dianthus caryophyllus	By Pagemoral - Contributor Pagemoral takes a photograph, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6745101	7
Silene latifolia	By Walter Siegmund (talk) - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=8730357	7
Cerastium arvense	By Walter Siegmund (talk) - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6626897	7
Pisonia umbellifera	By Forest & Kim Starr, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=6128755	8
Nyctaginia capitata	Michael J. Moore	8
Abronia umbellata	Michael J. Moore	8
Mirabilis multiflora	Michael J. Moore	8
Dionaea muscipula	Joseph F. Walker	4
Dionaea muscipula (flower)	Joseph F. Walker	4
Nepenthes alata	Joseph F. Walker	4
Drosophyllum lusitanicum	Joseph F. Walker	4
Fankenia laevis	By Ghislain118 http://www.fleurs-des-montagnes.net - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=12793355	4
Fagopyrum vesculentum	By Kurt Stüber [1] - caliban.mpiz-koeln.mpg.de/mavica/index.html part of www.biolib.de, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6268	4

Oxytheca perfoliata	By Stan Shebs, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=7845273	4

Table C-7 Photo credits for figure 4-8, including the author, license and where the picture was obtained from

Appendix D

Supplementary Figures and Tables for Chapter V

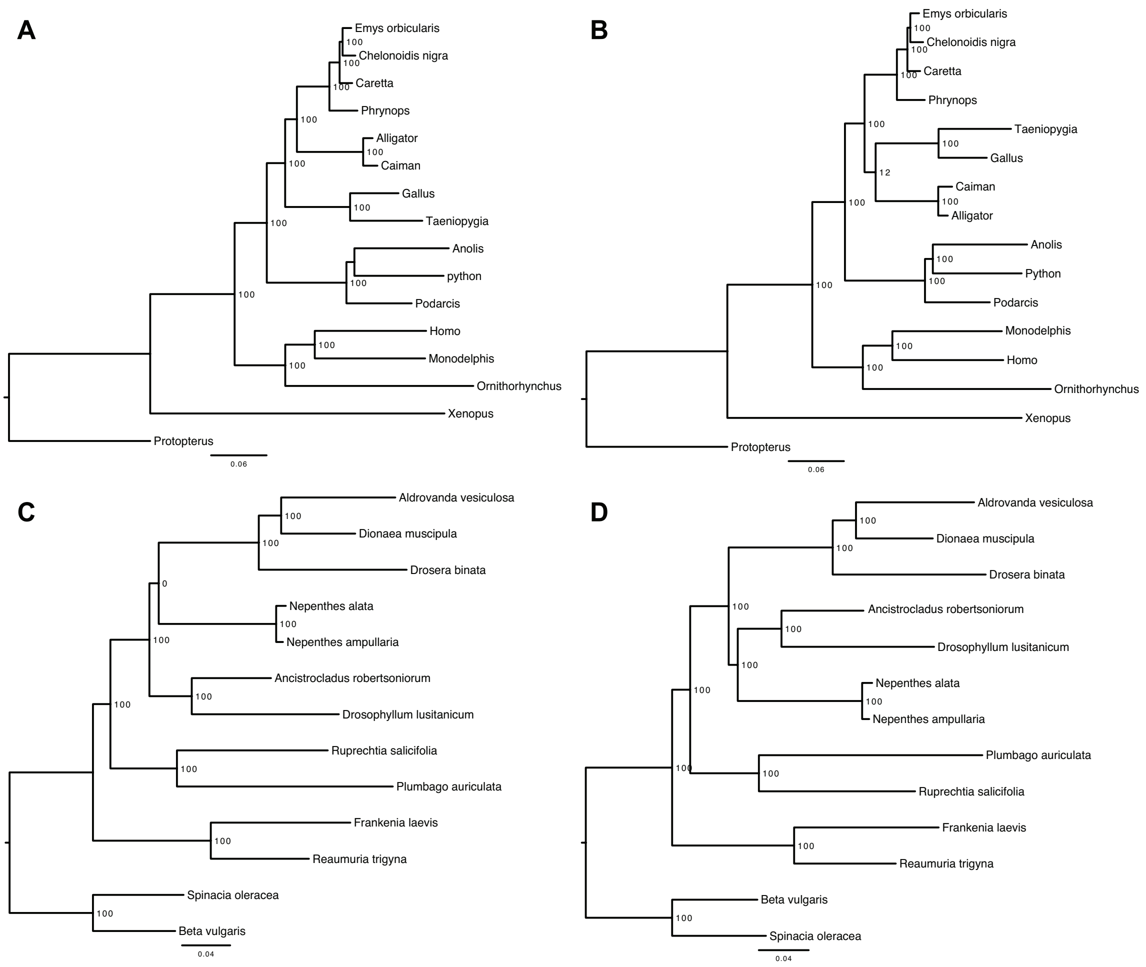


Figure D-1 Species trees inferred using maximum likelihood from the different supermatrices.

Support at each node was obtained from 200 rapid bootstrap replicates. A) Species tree for vertebrate dataset inferred with all 248 genes included in the supermatrix. B) Species tree for the vertebrate dataset inferred with 8916 and 11434 removed from the supermatrix. C) carnivorous Caryophyllales species tree inferred from all 1237 genes. D) carnivorous Caryophyllales species tree inferred with cluster575 and cluster3300 removed from the supermatrix.

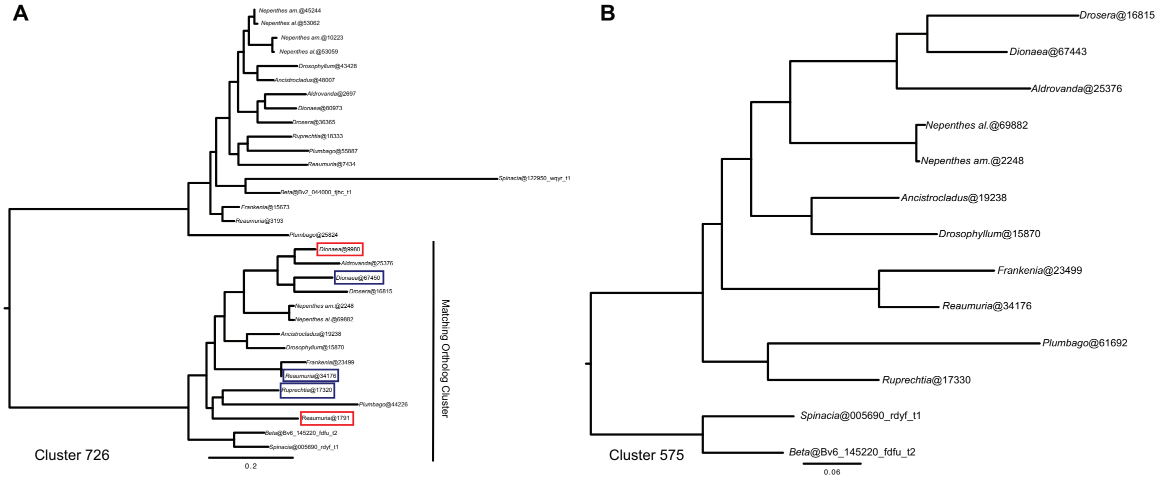


Figure D-2. Homolog tree for Amino Acid clustered (726) and CDS clustered (575) highly influential gene in the carnivorous Caryophyllales dataset.

Different genes identified in the ortholog clusters are circled on cluster 726. Genes circled in red represent ones that are shorter and were not identified as orthologous in the CDS dataset and genes circled in blue represent alternate paralogs or introsplice sites used between the two clustering analyses.

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

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

Ortholog in 575	Ortholog in 726	Seq length of 575 (Nuc)	Seq length of 726 (Nuc)	Reason for misidentification
Dino@67443 (*Dionaea*)	Dino@67450	2793	2991	Different copy of the in-paralog or intron splice site was retained
Dino@67443 (*Dionaea*)	Dino@9980	2793	510	Not identified as homologs in blast
RuprSFB@17320 (*Ruprechtia*)	RuprSFB@17330	2787	2787	Different copy of the in-paralog or intron splice site was retained
MJM3360@61692 (*Plumbago*)	MJM3360@44226	2211	2403	Different copy of the in-paralog or intron splice site was retained
Retr@34176 (*Reaumuria*)	Retr@1791	1044	546	Not identified as homologs in blast

Table D-2. Sources of discrepancy between the orthologs detected in highly influential nucleotide cluster575 and in matching amino acid homolog cluster726.

Bibliography

Adachi, Jun, Peter J. Waddell, William Martin, and Masami Hasegawa. "Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA." *Journal of Molecular Evolution* 50, no. 4 (2000): 348-358.

Albert, Victor A., Stephen E. Williams, and Mark W. Chase. "Carnivorous plants: phylogeny and structural evolution." *Science* 257, no. 5076 (1992): 1491-1495.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research* 25, no. 17 (1997): 3389-3402.

Ané, Cécile, Bret Larget, David A. Baum, Stacey D. Smith, and Antonis Rokas. "Bayesian estimation of concordance among gene trees." *Molecular biology and evolution* 24, no. 2 (2006): 412-426.

Anisimova, Maria, Manuel Gil, Jean-François Dufayard, Christophe Dessimoz, and Olivier Gascuel. "Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes." *Systematic biology* 60, no. 5 (2011): 685-699.

Arakaki, Mónica, Pascal-Antoine Christin, Reto Nyffeler, Anita Lendel, Urs Eggli, R. Matthew Ogburn, Elizabeth Spriggs, Michael J. Moore, and Erika J. Edwards. "Contemporaneous and recent radiations of the world's major succulent plant lineages." *Proceedings of the National Academy of Sciences*108, no. 20 (2011): 8379-8384.

Arcila, Dahiana, Guillermo Ortí, Richard Vari, Jonathan W. Armbruster, Melanie LJ Stiassny, Kyung D. Ko, Mark H. Sabaj, John Lundberg, Liam J. Revell, and Ricardo Betancur-R. "Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life." *Nature ecology & evolution* 1, no. 2 (2017): 0020.

Barker, Michael S., Nolan C. Kane, Marta Matvienko, Alexander Kozik, Richard W. Michelmore, Steven J. Knapp, and Loren H. Rieseberg. "Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years." *Molecular Biology and Evolution* 25, no. 11 (2008): 2445-2455.

Barker, Michael S., Zheng Li, Thomas I. Kidder, Chris R. Reardon, Zhao Lai, Luiz O. Oliveira, Moira Scascitelli, and Loren H. Rieseberg. "Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae." *American journal of botany* 103, no. 7 (2016): 1203-1211.

Bemm, Felix, Dirk Becker, Christina Larisch, Ines Kreuzer, Maria Escalante-Perez, Waltraud X. Schulze, Markus Ankenbrand et al. "Venus flytrap carnivorous lifestyle builds on herbivore defense strategies." *Genome research* 26, no. 6 (2016): 812-825.

Bendich, Arnold J. "Circular chloroplast chromosomes: the grand illusion." *The Plant Cell* 16, no. 7 (2004): 1661-1666.

Birky, C. William. "Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution." *Proceedings of the National Academy of Sciences* 92, no. 25 (1995): 11331-11338.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30, no. 15 (2014): 2114-2120.

Boussau, Bastien, Gergely J. Szöllősi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. "Genome-scale coestimation of species and gene trees." *Genome research*23, no. 2 (2013): 323-330.

Bradley, Robert K., Adam Roberts, Michael Smoot, Sudeep Juvekar, Jaeyoung Do, Colin Dewey, Ian Holmes, and Lior Pachter. "Fast statistical alignment." *PLoS computational biology* 5, no. 5 (2009): e1000392.

Brockington, Samuel F., Roolse Alexandre, Jeremy Ramdial, Michael J. Moore, Sunny Crawley, Amit Dhingra, Khidir Hilu, Douglas E. Soltis, and Pamela S. Soltis. "Phylogeny of the Caryophyllales sensu lato: revisiting hypotheses on pollination biology and perianth differentiation in the core Caryophyllales." *International Journal of Plant Sciences* 170, no. 5 (2009): 627-643.

Brockington, Samuel F., Rachel H. Walker, Beverley J. Glover, Pamela S. Soltis, and Douglas E. Soltis. "Complex pigment evolution in the Caryophyllales." *New Phytologist* 190, no. 4 (2011): 854-864.

Brockington, Samuel F., Ya Yang, Fernando Gandia-Herrero, Sarah Covshoff, Julian M. Hibberd, Rowan F. Sage, Gane KS Wong, Michael J. Moore, and Stephen A. Smith. "Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales." *New Phytologist* 207, no. 4 (2015): 1170-1180.

Brown, Jeremy M., and Robert C. Thomson. "Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses." *Systematic biology* 66, no. 4 (2016): 517-530.

Brown, Joseph W., Joseph F. Walker, and Stephen A. Smith. "Phyx: phylogenetic tools for unix." *Bioinformatics* 33, no. 12 (2017): 1886-1888.

A. Burleigh, J. Gordon, and Sarah Mathews. "Assessing among-locus variation in the inference of seed plant phylogeny." *International Journal of Plant Sciences* 168, no. 2 (2007): 111-124.

B. Burleigh, J. Gordon, and Sarah Mathews. "Assessing systematic error in the inference of seed plant phylogeny." *International Journal of Plant Sciences* 168, no. 2 (2007): 125-135.

Byng, James W., Mark W. Chase, Maarten JM Christenhusz, Michael F. Fay, Walter S. Judd, David J. Mabberley, Alexander N. Sennikov et al. "An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV." *Botanical Journal of the Linnean Society* 181, no. 1 (2016): 1-20.

Cameron, Kenneth M., Kenneth J. Wurdack, and Richard W. Jobson. "Molecular evidence for the common origin of snap-traps among carnivorous plants." *American Journal of Botany*89, no. 9 (2002): 1503-1509.

Camin, Joseph H., and Robert R. Sokal. "A method for deducing branching sequences in phylogeny." *Evolution* 19, no. 3 (1965): 311-326.

Cannon, Steven B., Michael R. McKain, Alex Harkess, Matthew N. Nelson, Sudhansu Dash, Michael K. Deyholos, Yanhui Peng et al. "Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes." *Molecular biology and evolution* 32, no. 1 (2014): 193-210.

Castoe, Todd A., AP Jason de Koning, Hyun-Min Kim, Wanjun Gu, Brice P. Noonan, Gavin Naylor, Zhi J. Jiang, Christopher L. Parkinson, and David D. Pollock. "Evidence for an ancient adaptive episode of convergent molecular evolution." *Proceedings of the National Academy of Sciences* 106, no. 22 (2009): 8986-8991.

Cavieres, Lohengrin A., Patricia Sáez, Carolina Sanhueza, Angela Sierra-Almeida, Claudia Rabert, Luis J. Corcuera, Miren Alberdi, and León A. Bravo. "Ecophysiological traits of Antarctic vascular plants: their importance in the responses to climate change." *Plant ecology* 217, no. 3 (2016): 343-358.

Chase, Mark W., Douglas E. Soltis, Richard G. Olmstead, David Morgan, Donald H. Les, Brent D. Mishler, Melvin R. Duvall et al. "Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL." *Annals of the Missouri Botanical Garden* (1993): 528-580.

Cheng, Shifeng, Michael Melkonian, Stephen A. Smith, Samuel Brockington, John M. Archibald, Pierre-Marc Delaux, Fay-Wei Li et al. "10KP: A phylodiverse genome sequencing plan." *GigaScience* 7, no. 3 (2018): giy013.

Chiari, Ylenia, Vincent Cahais, Nicolas Galtier, and Frédéric Delsuc. "Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria)." *Bmc Biology* 10, no. 1 (2012): 65.

Christenhusz, Maarten JM, Samuel F. Brockington, Pascal-Antoine Christin, and Rowan F. Sage. "On the disintegration of Molluginaceae: a new genus and family (Kewa, Kewaceae) segregated from Hypertelis, and placement of Macarthuria in Macarthuriaceae." *Phytotaxa* 181, no. 4 (2014): 238-242.

Christin, Pascal-Antoine, Erika J. Edwards, Guillaume Besnard, Susanna F. Boxall, Richard Gregory, Elizabeth A. Kellogg, James Hartwell, and Colin P. Osborne. "Adaptive evolution of C 4 photosynthesis through recurrent lateral gene transfer." *Current Biology* 22, no. 5 (2012): 445-449.

Christin, Pascal-Antoine, Tammy L. Sage, Erika J. Edwards, R. Matthew Ogburn, Roxana Khoshravesh, and Rowan F. Sage. "Complex evolutionary transitions and the significance of C3–C4 intermediate forms of photosynthesis in Molluginaceae." *Evolution* 65, no. 3 (2011): 643-660.

Cronn, Richard, Aaron Liston, Matthew Parks, David S. Gernandt, Rongkun Shen, and Todd Mockler. "Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology." *Nucleic acids research*36, no. 19 (2008): e122-e122.

Cronn, Richard, Brian J. Knaus, Aaron Liston, Peter J. Maughan, Matthew Parks, John V. Syring, and Joshua Udall. "Targeted enrichment strategies for next-generation plant biology." *American Journal of Botany* 99, no. 2 (2012): 291-311.

Cuénoud, Philippe, Vincent Savolainen, Lars W. Chatrou, Martyn Powell, Renée J. Grayer, and Mark W. Chase. "Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences." *American Journal of Botany* 89, no. 1 (2002): 132-144.

Dang, Zhen-hua, Lin-lin Zheng, Jia Wang, Zhe Gao, Shu-biao Wu, Zhi Qi, and Ying-chun Wang. "Transcriptomic profiling of the salt-stress response in the wild recretohalophyte Reaumuria trigyna." *BMC genomics* 14, no. 1 (2013): 29.

Darwin, C., 1859. *The origin of species by means of natural selection*. Modern Lib..

Darwin, Charles, and Francis Darwin. *Insectivorous plants*. J. Murray, 1888.

Dasmahapatra, Kanchon K., James R. Walters, Adriana D. Briscoe, John W. Davey, Annabel Whibley, Nicola J. Nadeau, Aleksey V. Zimin et al. "Butterfly genome reveals promiscuous exchange of mimicry adaptations among species." *Nature* 487, no. 7405 (2012): 94.

Degreef, John D. "Fossil Aldrovanda." *Carnivorous Plant Newsletter* 26 (1997): 93-97.

Dohm, Juliane C., Cornelia Lange, Daniela Holtgräwe, Thomas Rosleff Sörensen, Dietrich Borchardt, Britta Schulz, Hans Lehrach, Bernd Weisshaar, and Heinz Himmelbauer. "Palaeohexaploid ancestry for Caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (Beta vulgaris)." *The plant journal* 70, no. 3 (2012): 528-540.

Dohm, Juliane C., André E. Minoche, Daniela Holtgräwe, Salvador Capella-Gutiérrez, Falk Zakrzewski, Hakim Tafer, Oliver Rupp et al. "The genome of the recently domesticated crop plant sugar beet (Beta vulgaris)." *Nature* 505, no. 7484 (2014): 546.

Dong, Wenpan, Chao Xu, Changhao Li, Jiahui Sun, Yunjuan Zuo, Shuo Shi, Tao Cheng, Junjie Guo, and Shiliang Zhou. "ycf1, the most promising plastid DNA barcode of land plants." *Scientific reports* 5 (2015): 8348.

Donoghue, Michael J., Richard G. Olmstead, James F. Smith, and Jeffrey D. Palmer. "Phylogenetic relationships of Dipsacales based on rbcL sequences." *Annals of the Missouri Botanical Garden* (1992): 333-345.

Drummond, Alexei J., and Andrew Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees." *BMC evolutionary biology* 7, no. 1 (2007): 214.

Dunn, Casey W., Andreas Hejnol, David Q. Matus, Kevin Pang, William E. Browne, Stephen A. Smith, Elaine Seaver et al. "Broad phylogenomic sampling improves resolution of the animal tree of life." *Nature* 452, no. 7188 (2008): 745.

Dunn, Casey W., Mark Howison, and Felipe Zapata. "Agalma: an automated phylogenomics workflow." *BMC bioinformatics*14, no. 1 (2013): 330.

Eaton, Deren AR, Elizabeth L. Spriggs, Brian Park, and Michael J. Donoghue. "Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants." *Systematic biology* 66, no. 3 (2017): 399-412.

Edger, Patrick P., Hanna M. Heidel-Fischer, Michaël Bekaert, Jadranka Rota, Gernot Glöckner, Adrian E. Platts, David G. Heckel et al. "The butterfly plant arms-race escalated by gene and genome duplications." *Proceedings of the National Academy of Sciences* 112, no. 27 (2015): 8362-8366.

Eggli, Urs, and Reto Nyffeler. "Living under temporarily arid conditions-succulence as an adaptive strategy." *Bradleya* 27 (2009): 13-36.

Ellison, Aaron M., and Nicholas J. Gotelli. "Energetics and the evolution of carnivorous plants—Darwin's 'most wonderful plants in the world'." *Journal of Experimental Botany* 60, no. 1 (2009): 19-42.

Ellison, Aaron M., and Nicholas J. Gotelli. "Evolutionary ecology of carnivorous plants." *Trends in ecology & evolution*16, no. 11 (2001): 623-629.

Emms, David M., and Steven Kelly. "OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy." *Genome biology* 16, no. 1 (2015): 157.

Edwards, Erika J., Reto Nyffeler, and Michael J. Donoghue. "Basal cactus phylogeny: implications of Pereskia (Cactaceae) paraphyly for the transition to the cactus life form." *American Journal of Botany* 92, no. 7 (2005): 1177-1188.

Felsenstein, Joseph. "Evolutionary trees from DNA sequences: a maximum likelihood approach." *Journal of molecular evolution* 17, no. 6 (1981): 368-376.

Felsenstein, Joseph. "Statistical inference of phylogenies." *Journal of the Royal Statistical Society. Series A (General)*(1983): 246-272.

Felsenstein, Joseph. "Confidence limits on phylogenies: an approach using the bootstrap." *Evolution* 39, no. 4 (1985): 783-791.

Fernández-Mazuecos, Mario, Greg Mellers, Beatriz Vigalondo, Llorenç Sáez, Pablo Vargas, and Beverley J. Glover. "Resolving Recent Plant Radiations: Power and Robustness of Genotyping-by-Sequencing." *Systematic biology* (2017).

Flowers, Timothy J., and Timothy D. Colmer. "Salinity tolerance in halophytes." *New Phytologist* 179, no. 4 (2008): 945-963.

Fontaine, Michael C., James B. Pease, Aaron Steele, Robert M. Waterhouse, Daniel E. Neafsey, Igor V. Sharakhov, Xiaofang Jiang et al. "Extensive introgression in a malaria vector species complex revealed by phylogenomics." *Science*347, no. 6217 (2015): 1258524.

Foster, Peter G. "Modeling compositional heterogeneity." *Systematic Biology* 53, no. 3 (2004): 485-495.

Freeling, Michael, and Brian C. Thomas. "Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity." *Genome research* 16, no. 7 (2006): 805-814.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. "CD-HIT: accelerated for clustering the next-generation sequencing data." *Bioinformatics* 28, no. 23 (2012): 3150-3152.

Fu, Qiaomei, Mateja Hajdinjak, Oana Teodora Moldovan, Silviu Constantin, Swapan Mallick, Pontus Skoglund, Nick Patterson et al. "An early modern human from Romania with a recent Neanderthal ancestor." *Nature* 524, no. 7564 (2015): 216.

Gabaldón, Toni. "Large-scale assignment of orthology: back to phylogenetics?." *Genome biology* 9, no. 10 (2008): 235.

Galtier, Nicolas, and Vincent Daubin. "Dealing with incongruence in phylogenomic analyses." *Philosophical Transactions of the Royal Society B: Biological Sciences* 363, no. 1512 (2008): 4023-4029.

Gitzendanner, Matthew A., Pamela S. Soltis, Gane K-S. Wong, Brad R. Ruhfel, and Douglas E. Soltis. "Plastid phylogenomic analysis of green plants: a billion years of evolutionary history." *American journal of botany* (2018).

Givnish, Thomas J. "New evidence on the origin of carnivorous plants." *Proceedings of the National Academy of Sciences* 112, no. 1 (2015): 10-11.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis et al. "Full-length transcriptome assembly from RNA-Seq data without a reference genome." *Nature biotechnology* 29, no. 7 (2011): 644.

Graham, Sean W., and Richard G. Olmstead. "Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms." *American Journal of Botany* 87, no. 11 (2000): 1712-1730.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0." *Systematic biology* 59, no. 3 (2010): 307-321.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger et al. "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." *Nature protocols* 8, no. 8 (2013): 1494.

Haeckel, Ernst. *bd. Allgemeine entwickelungsgeschichte der organismen*. Vol. 2. G. Reimer, 1866.

Harkess, Alex, Jinsong Zhou, Chunyan Xu, John E. Bowers, Ron Hulst, Saravanaraj Ayyampalayam, Francesco Mercati et al. "The asparagus genome sheds light on the origin and evolution of a young Y chromosome." *Nature Communications*8, no. 1 (2017): 1279.

Hennig, W., 1965. Phylogenetic systematics. *Annual review of entomology*, *10*(1), pp.97-116.

Hernández-Ledesma, Patricia, Walter G. Berendsohn, Thomas Borsch, Sabine Von Mering, Hossein Akhani, Salvador Arias, Idelfonso Castañeda-Noa et al. "A taxonomic backbone for the global synthesis of species diversity in the angiosperm order Caryophyllales." *Willdenowia* 45, no. 3 (2015): 281-383.

Heubl, G., G. Bringmann, and H. Meimberg. "Molecular phylogeny and character evolution of carnivorous plant families in Caryophyllales—revisited." *Plant Biology* 8, no. 6 (2006): 821-830.

Hilu, Khidir W., Thomas Borsch, Kai Müller, Douglas E. Soltis, Pamela S. Soltis, Vincent Savolainen, Mark W. Chase et al. "Angiosperm phylogeny based on< 011> matK sequence information." *American journal of botany* 90, no. 12 (2003): 1758-1776.

Hlavac, Marek. "stargazer: LaTeX code and ASCII text for well-formatted regression and summary statistics tables." URL: http://CRAN. R-project. org/package= stargazer (2013).

Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Sy Vinh Le. "UFBoot2: Improving the Ultrafast Bootstrap Approximation." *Molecular biology and evolution* (2017): msx281.

Jansen, Robert K., Zhengqiu Cai, Linda A. Raubeson, Henry Daniell, James Leebens-Mack, Kai F. Müller, Mary Guisinger-Bellian et al. "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, no. 49 (2007): 19369-19374.

Jarvis, Erich D., Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon YW Ho et al. "Whole-genome analyses resolve early branches in the tree of life of modern birds." *Science* 346, no. 6215 (2014): 1320-1331.

Jian, Shuguang, Pamela S. Soltis, Matthew A. Gitzendanner, Michael J. Moore, Ruiqi Li, Tory A. Hendry, Yin-Long Qiu, Amit Dhingra, Charles D. Bell, and Douglas E. Soltis. "Resolving an ancient, rapid radiation in Saxifragales." *Systematic Biology*57, no. 1 (2008): 38-57.

Jukes, Thomas H., and Charles R. Cantor. "Evolution of protein molecules." *Mammalian protein metabolism* 3, no. 21 (1969): 132.

Kadereit, Gudrun, David Ackerly, and Michael D. Pirie. "A broader model for C4 photosynthesis evolution in plants inferred from the goosefoot family (Chenopodiaceae ss)." *Proc. R. Soc. B* 279, no. 1741 (2012): 3304-3311.

Kadereit, G., Th Borsch, K. Weising, and H. Freitag. "Phylogeny of Amaranthaceae and Chenopodiaceae and the evolution of C4 photosynthesis." *International journal of plant sciences* 164, no. 6 (2003): 959-986.

Kainer, David, and Robert Lanfear. "The effects of partitioning on phylogenetic inference." *Molecular biology and evolution*32, no. 6 (2015): 1611-1627.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas KF Wong, Arndt von Haeseler, and Lars S. Jermiin. "ModelFinder: fast model selection for accurate phylogenetic estimates." *Nature methods* 14, no. 6 (2017): 587.

Katoh, Kazutaka, and Daron M. Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." *Molecular biology and evolution*30, no. 4 (2013): 772-780.

Kelly, Steven, Sarah Covshoff, Samart Wanchana, Vivek Thakur, Paul Quick, Yu Wang, Martha Ludwig et al. "Wide sampling of natural diversity identifies novel molecular signatures of C4 photosynthesis." *bioRxiv* (2017): 163097.

Kobert, Kassian, Leonidas Salichos, Antonis Rokas, and Alexandros Stamatakis. "Computing the internode certainty and related measures from partial gene trees." *Molecular biology and evolution* 33, no. 6 (2016): 1606-1617.

Kocot, Kevin M., Mathew R. Citarella, Leonid L. Moroz, and Kenneth M. Halanych. "PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics." *Evolutionary Bioinformatics* 9 (2013): EBO-S12813.

Lanfear, Robert, Brett Calcott, Simon YW Ho, and Stephane Guindon. "PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses." *Molecular biology and evolution* 29, no. 6 (2012): 1695-1701.

Larkin, Mark A., Gordon Blackshields, N. P. Brown, R. Chenna, Paul A. McGettigan, Hamish McWilliam, Franck Valentin et al. "Clustal W and Clustal X version 2.0." *bioinformatics* 23, no. 21 (2007): 2947-2948.

Lee, M. S. Y., and A. F. Hugall. "Partitioned likelihood support and the evaluation of data set conflict." *Systematic biology* 52, no. 1 (2003): 15-22.

Leigh, Jessica W., Edward Susko, Manuela Baumgartner, and Andrew J. Roger. "Testing congruence in phylogenomic analysis." *Systematic Biology* 57, no. 1 (2008): 104-115.

Liu, Juan, Zhe-Chen Qi, Yun-Peng Zhao, Cheng-Xin Fu, and Qiu-Yun Jenny Xiang. "Complete cpDNA genome sequence of Smilax china and phylogenetic placement of Liliales–Influences of gene partitions and taxon sampling." *Molecular phylogenetics and evolution* 64, no. 3 (2012): 545-562.

Lopez-Nieves, Samuel, Ya Yang, Alfonso Timoneda, Minmin Wang, Tao Feng, Stephen A. Smith, Samuel F. Brockington, and Hiroshi A. Maeda. "Relaxation of tyrosine pathway regulation underlies the evolution of betalain pigmentation in Caryophyllales." *New Phytologist* 217, no. 2 (2018): 896-908.

Löytynoja, Ari, and Nick Goldman. "Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis." *Science* 320, no. 5883 (2008): 1632-1635.

Maddison, Wayne P. "Gene trees in species trees." *Systematic biology* 46, no. 3 (1997): 523-536.

Magallón, Susana, Sandra Gómez-Acevedo, Luna L. Sánchez-Reyes, and Tania Hernández-Hernández. "A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity." *New Phytologist* 207, no. 2 (2015): 437-453.

Martin, Andrew P., and Theresa M. Burg. "Perils of paralogy: using HSP70 genes for inferring organismal phylogenies." *Systematic Biology* 51, no. 4 (2002): 570-587.

Martin, William. "Gene transfer from organelles to the nucleus: frequent and in big chunks." *Proceedings of the National Academy of Sciences* 100, no. 15 (2003): 8612-8614.

Martin, William, Bettina Stoebe, Vadim Goremykin, Sabine Hansmann, Masami Hasegawa, and Klaus V. Kowallik. "Gene transfer to the nucleus and the evolution of chloroplasts." *Nature* 393, no. 6681 (1998): 162.

Masson, Rüdiger, and Gudrun Kadereit. "Phylogeny of Polycnemoideae (Amaranthaceae): Implications for biogeography, character evolution and taxonomy." *Taxon* 62, no. 1 (2013): 100-111.

Mayrose, Itay, Shing H. Zhan, Carl J. Rothfels, Karen Magnuson-Ford, Michael S. Barker, Loren H. Rieseberg, and Sarah P. Otto. "Recently formed polyploid plants diversify at lower rates." *Science* 333, no. 6047 (2011): 1257-1257.

McCauley, David E., Allyson K. Sundby, Maia F. Bailey, and Mark E. Welch. "Inheritance of chloroplast DNA is not strictly maternal in Silene vulgaris (Caryophyllaceae): evidence from experimental crosses and natural populations." *American Journal of Botany* 94, no. 8 (2007): 1333-1337.

McKain, Michael R., Haibao Tang, Joel R. McNeal, Saravanaraj Ayyampalayam, Jerrold I. Davis, Claude W. dePamphilis, Thomas J. Givnish, J. Chris Pires, Dennis Wm Stevenson, and James H. Leebens-Mack. "A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales." *Genome biology and evolution* 8, no. 4 (2016): 1150-1164.

McPherson, Stewart, and Alastair Robinson. *Pitcher plants of the Old World*. Vol. 2. Poole: Redfern Natural History Productions, 2009.
Meimberg, H., P. Dittrich, G. Bringmann, J. Schlauer, and G. Heubl. "Molecular phylogeny of Caryophyllidae sl based on matK sequences with special emphasis on carnivorous taxa." *Plant Biology* 2, no. 2 (2000): 218-228.

Mendes, Fabio K., and Matthew W. Hahn. "Gene tree discordance causes apparent substitution rate variation." *Systematic biology* 65, no. 4 (2016): 711-721.

Mirarab, Siavash, Rezwana Reaz, Md S. Bayzid, Théo Zimmermann, M. Shel Swenson, and Tandy Warnow. "ASTRAL: genome-scale coalescent-based species tree estimation." *Bioinformatics* 30, no. 17 (2014): i541-i548.

Moore, Abigail J., Jurriaan M. De Vos, Lillian P. Hancock, Eric Goolsby, and Erika J. Edwards. "Targeted Enrichment of Large Gene Families for Phylogenetic Inference: Phylogeny and Molecular Evolution of Photosynthesis Genes in the Portullugo Clade (Caryophyllales)." *Systematic biology* (2017).

Moore, Michael J., Amit Dhingra, Pamela S. Soltis, Regina Shaw, William G. Farmerie, Kevin M. Folta, and Douglas E. Soltis. "Rapid and accurate pyrosequencing of angiosperm plastid genomes." *BMC Plant Biology* 6, no. 1 (2006): 17.

Moore, Michael J., Charles D. Bell, Pamela S. Soltis, and Douglas E. Soltis. "Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms." *Proceedings of the National Academy of Sciences* 104, no. 49 (2007): 19363-19368.

Moore, Michael J., Pamela S. Soltis, Charles D. Bell, J. Gordon Burleigh, and Douglas E. Soltis. "Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots." *Proceedings of the National Academy of Sciences* 107, no. 10 (2010): 4623-4628.

Moore, Michael J., Nasr Hassan, Matthew A. Gitzendanner, Riva A. Bruenn, Matthew Croley, Alexia Vandeventer, James W. Horn et al. "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, no. 4 (2011): 541-558.

Moray, Camile, Eric W. Goolsby, and Lindell Bromham. "The phylogenetic association between salt tolerance and heavy metal hyperaccumulation in angiosperms." *Evolutionary Biology* 43, no. 1 (2016): 119-130.

Morgan, Claire C., Peter G. Foster, Andrew E. Webb, Davide Pisani, James O. McInerney, and Mary J. O'Connell. "Heterogeneous models place the root of the placental mammal phylogeny." *Molecular biology and evolution* 30, no. 9 (2013): 2145-2156.

Nei, Masatoshi, and Takashi Gojobori. "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." *Molecular biology and evolution* 3, no. 5 (1986): 418-426.

Neubig, Kurt M., and J. Richard Abbott. "Primer development for the plastid region ycf1 in Annonaceae and other magnoliids." *American journal of botany* 97, no. 6 (2010).

Neubig, Kurt M., W. Mark Whitten, Barbara S. Carlsward, Mario A. Blanco, Lorena Endara, Norris H. Williams, and Michael Moore. "Phylogenetic utility of ycf1 in orchids: a plastid gene more variable than matK." *Plant Systematics and Evolution* 277, no. 1-2 (2009): 75-84.

Neupane, Suman, Karolina Fucikova, Louise A. Lewis, Lynn Kuo, Ming-Hui Chen, and Paul Lewis. "Assessing Combinability of Phylogenomic Data using Bayes Factors." *bioRxiv* (2018): 250969.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies." *Molecular biology and evolution* 32, no. 1 (2014): 268-274.

Ohno, Susumu, Ulrich Wolf, and Niels B. Atkin. "Evolution from fish to mammals by gene duplication." *Hereditas* 59, no. 1 (1968): 169-187.

Olave, Melisa, Luciano J. Avila, Jack W. Sites Jr, and Mariana Morando. "Model-based approach to test hard polytomies in the Eulaemus clade of the most diverse South American lizard genus Liolaemus (Liolaemini, Squamata)." *Zoological Journal of the Linnean Society* 174, no. 1 (2015): 169-184.

Palmer, Jeffrey D. "Chloroplast DNA exists in two orientations." *Nature* 301, no. 5895 (1983): 92.

Palmer, Jeffrey D. "Comparative organization of chloroplast genomes." *Annual review of genetics* 19, no. 1 (1985): 325-354.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. "APE: analyses of phylogenetics and evolution in R language." Bioinformatics 20, no. 2 (2004): 289-290.

Pease, James B., David C. Haak, Matthew W. Hahn, and Leonie C. Moyle. "Phylogenomics reveals three sources of adaptive variation during a rapid radiation." *PLoS Biology* 14, no. 2 (2016): e1002379.

Pease, James B., Joseph W. Brown, Joseph F. Walker, Cody E. Hinchliff, and Stephen A. Smith. "Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life." *American journal of botany* 105, no. 3 (2018): 385-403.

Philippe, Hervé, and Patrick Forterre. "The rooting of the universal tree of life is not reliable." Journal of Molecular Evolution 49, no. 4 (1999): 509-523.

Philippe, Hervé, Henner Brinkmann, Dennis V. Lavrov, D. Timothy J. Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. "Resolving difficult phylogenetic questions: why more sequences are not enough." *PLoS biology* 9, no. 3 (2011): e1000602.

Posada, David, and Thomas R. Buckley. "Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests." *Systematic biology* 53, no. 5 (2004): 793-808.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. "FastTree 2–approximately maximum-likelihood trees for large alignments." *PloS one* 5, no. 3 (2010): e9490.

Prum, Richard O., Jacob S. Berv, Alex Dornburg, Daniel J. Field, Jeffrey P. Townsend, Emily Moriarty Lemmon, and Alan R. Lemmon. "A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing." *Nature* 526, no. 7574 (2015): 569.

Redelings, Benjamin. "Erasing errors due to alignment ambiguity when estimating positive selection." *Molecular biology and evolution* 31, no. 8 (2014): 1979-1993.

Rettig, J. H., Hugh D. Wilson, and James R. Manhart. "Phylogeny of the Caryophyllales: Gene sequence data." *Taxon* (1992): 201-209.

Richards, Emilie J., Jeremy M. Brown, Anthony J. Barley, Rebecca A. Chong, and Robert C. Thomson. "Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological?." *Systematic biology* (2018): syy013.

Rodman, James E., Michael K. Oliver, Robert R. Nakamura, James U. McClammer Jr, and Anthony H. Bledsoe. "A taxonomic analysis and revised classification of Centrospermae." *Systematic Botany* (1984): 297-323.

Rodríguez-Ezpeleta, Naiara, Henner Brinkmann, Gertraud Burger, Andrew J. Roger, Michael W. Gray, Hervé Philippe, and B. Franz Lang. "Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans." *Current Biology* 17, no. 16 (2007): 1420-1425.

Rokas, Antonis, Barry L. Williams, Nicole King, and Sean B. Carroll. "Genome-scale approaches to resolving incongruence in molecular phylogenies." *Nature* 425, no. 6960 (2003): 798.

Romiguier, Jonathan, Vincent Ranwez, Frédéric Delsuc, Nicolas Galtier, and Emmanuel JP Douzery. "Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals." *Molecular biology and evolution* 30, no. 9 (2013): 2134-2144.

Roquet, Cristina, Isabel Sanmartín, Núria Garcia-Jacas, Llorenç Sáez, Alfonso Susanna, Niklas Wikström, and Juan José Aldasoro. "Reconstructing the history of Campanulaceae with a Bayesian approach to molecular dating and dispersal–vicariance analyses." *Molecular Phylogenetics and Evolution*52, no. 3 (2009): 575-587.

Ruhfel, Brad R., Matthew A. Gitzendanner, Pamela S. Soltis, Douglas E. Soltis, and J. Gordon Burleigh. "From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes." *BMC Evolutionary Biology* 14, no. 1 (2014): 23.

Ryan, Joseph F., Kevin Pang, Christine E. Schnitzler, Anh-Dao Nguyen, R. Travis Moreland, David K. Simmons, Bernard J. Koch et al. "The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution." *Science* 342, no. 6164 (2013): 1242592.

Sage, Rowan F. "A portrait of the C4 photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineages, and Hall of Fame." *Journal of experimental botany* 68, no. 2 (2016): e11-e28.

Sage, Rowan F., Pascal-Antoine Christin, and Erika J. Edwards. "The C4 plant lineages of planet Earth." *Journal of Experimental Botany* 62, no. 9 (2011): 3155-3169.

Saitou, Naruya, and Masatoshi Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular biology and evolution* 4, no. 4 (1987): 406-425.

Sajeva, M., and J. D. Mauseth. "Leaf-like structure in the photosynthetic, succulent stems of cacti." *Annals of Botany*68, no. 5 (1991): 405-411.

Salichos, Leonidas, Alexandros Stamatakis, and Antonis Rokas. "Novel information theory-based measures for quantifying incongruence among phylogenetic trees." *Molecular Biology and Evolution* 31, no. 5 (2014): 1261-1271.

Sancho, Rubén, Carlos P. Cantalapiedra, Diana López-Alvarez, Sean P. Gordon, John P. Vogel, Pilar Catalán, and Bruno Contreras-Moreira. "Comparative plastome genomics and phylogenomics of Brachypodium: flowering time signatures, introgression and recombination in recently diverged ecotypes." *New Phytologist* 218, no. 4 (2018): 1631-1644.

Sayyari, Erfan, and Siavash Mirarab. "Fast coalescent-based computation of local branch support from quartet frequencies." *Molecular biology and evolution* 33, no. 7 (2016): 1654-1668.

Schäferhoff, Bastian, Kai F. Müller, and Thomas Borsch. "Caryophyllales phylogenetics: disentangling Phytolaccaceae and Molluginaceae and description of Microteaceae as a new isolated family." *Willdenowia* 39, no. 2 (2010): 209-228.

Schlueter, Jessica A., Phillip Dixon, Cheryl Granger, David Grant, Lynn Clark, Jeff J. Doyle, and Randy C. Shoemaker. "Mining EST databases to resolve evolutionary events in major crop species." *Genome* 47, no. 5 (2004): 868-876.

Seo, Tae-Kun. "Calculating bootstrap probabilities of phylogeny using multilocus sequence data." *Molecular biology and evolution* 25, no. 5 (2008): 960-971.

Shaw, Joey, Edgar B. Lickey, John T. Beck, Susan B. Farmer, Wusheng Liu, Jermey Miller, Kunsiri C. Siripun, Charles T. Winder, Edward E. Schilling, and Randall L. Small. "The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis." *American journal of botany* 92, no. 1 (2005): 142-166.

Shaw, Joey, Edgar B. Lickey, Edward E. Schilling, and Randall L. Small. "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, no. 3 (2007): 275-288.

Shaw, Joey, Hayden L. Shafer, O. Rayne Leonard, Margaret J. Kovach, Mark Schorr, and Ashley B. Morris. "Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV." *American Journal of Botany* 101, no. 11 (2014): 1987-2004.

Shen, Xing-Xing, Chris Todd Hittinger, and Antonis Rokas. "Contentious relationships in phylogenomic studies can be driven by a handful of genes." *Nature ecology & evolution* 1, no. 5 (2017): 0126.

Shepherd, Lara D., Barbara R. Holland, and Leon R. Perrie. "Conflict amongst chloroplast DNA sequences obscures the phylogeny of a group of Asplenium ferns." *Molecular Phylogenetics and Evolution* 48, no. 1 (2008): 176-187.

Smith, Stephen A., and Joseph W. Brown. "Constructing a broadly inclusive seed plant phylogeny." *American journal of botany* 105, no. 3 (2018): 302-314.

Smith, Stephen A., Joseph W. Brown, and Joseph F. Walker. "So many genes, so little time: A practical approach to divergence-time estimation in the genomic era." *PloS one* 13, no. 5 (2018): e0197433.

Smith, Stephen A., Joseph W. Brown, Ya Yang, Riva Bruenn, Chloe P. Drummond, Samuel F. Brockington, Joseph F. Walker, Noah Last, Norman A. Douglas, and Michael J. Moore. "Disparity, diversity, and duplications in the Caryophyllales." *New Phytologist* 217, no. 2 (2018): 836-854.

Smith, Stephen A., and Michael J. Donoghue. "Rates of molecular evolution are linked to life history in flowering plants." *science* 322, no. 5898 (2008): 86-89.

Smith, Stephen A., and Casey W. Dunn. "Phyutility: a phyloinformatics tool for trees, alignments and molecular data." *Bioinformatics* 24, no. 5 (2008): 715-716.

Smith, Stephen A., Michael J. Moore, Joseph W. Brown, and Ya Yang. "Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants." *BMC evolutionary biology* 15, no. 1 (2015): 150.

Smith, Stephen A., and James B. Pease. "Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny." *Briefings in bioinformatics* 18, no. 3 (2017): 451-457.

Smith, Stephen A., Nerida G. Wilson, Freya E. Goetz, Caitlin Feehery, Sónia CS Andrade, Greg W. Rouse, Gonzalo Giribet, and Casey W. Dunn. "Resolving the evolutionary relationships of molluscs with phylogenomic tools." *Nature* 480, no. 7377 (2011): 364.

Sokal, Robert R. "A statistical method for evaluating systematic relationship." *University of Kansas science bulletin*28 (1958): 1409-1438.

Soltis, Douglas E., Stephen A. Smith, Nico Cellinese, Kenneth J. Wurdack, David C. Tank, Samuel F. Brockington, Nancy F. Refulio-Rodriguez et al. "Angiosperm phylogeny: 17 genes, 640 taxa." *American journal of botany* 98, no. 4 (2011): 704-730.

Soltis, Douglas E., Pamela S. Soltis, Mark W. Chase, Mark E. Mort, Dirk C. Albach, Michael Zanis, Vincent Savolainen. "Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences." *Botanical Journal of the Linnean Society* 133, no. 4 (2000): 381-461.

Soltis, Douglas E., Stephen A. Smith, Nico Cellinese, Kenneth J. Wurdack, David C. Tank, Samuel F. Brockington, Nancy F. Refulio-Rodriguez et al. "Angiosperm phylogeny: 17 genes, 640 taxa." *American journal of botany* 98, no. 4 (2011): 704-730.

Soltis, Pam, Doug Soltis, and Monica Arakaki. "non-core Caryophyllales."

Soltis, Pamela S., Xiaoxian Liu, D. Blaine Marchant, Clayton J. Visger, and Douglas E. Soltis. "Polyploidy and novelty: Gottlieb's legacy." *Phil. Trans. R. Soc. B* 369, no. 1648 (2014): 20130351.

Stamatakis, Alexandros. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* 30, no. 9 (2014): 1312-1313.

Stamatakis, Alexandros, and Nikolaos Alachiotis. "Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data." *Bioinformatics*26, no. 12 (2010): i132-i139.

Stevens, Peter Francis, and Hilary Davis. "Angiosperm phylogeny website." (2001).

Stull, Gregory W., Michael J. Moore, Venkata S. Mandala, Norman A. Douglas, Heather-Rose Kates, Xinshuai Qi, Samuel F. Brockington, Pamela S. Soltis, Douglas E. Soltis, and Matthew A. Gitzendanner. "A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes." *Applications in Plant Sciences* 1, no. 2 (2013): 1200497.

Stull, Gregory W., Rodrigo Duno de Stefano, Douglas E. Soltis, and Pamela S. Soltis. "Resolving basal lamiid phylogeny and the circumscription of Icacinaceae with a plastome-scale data set." *American journal of botany* 102, no. 11 (2015): 1794-1813.

Suchard, Marc A., and Benjamin D. Redelings. "BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny." *Bioinformatics* 22, no. 16 (2006): 2047-2048.

Sullivan, Alexis R., Bastian Schiffthaler, Stacey Lee Thompson, Nathaniel R. Street, and Xiao-Ru Wang. "Interspecific plastome recombination reflects ancient reticulate evolution in Picea (Pinaceae)." *Molecular biology and evolution* 34, no. 7 (2017): 1689-1701.

Taberlet, Pierre, Ludovic Gielly, Guy Pautou, and Jean Bouvet. "Universal primers for amplification of three non-coding regions of chloroplast DNA." *Plant molecular biology* 17, no. 5 (1991): 1105-1109.

Tank, David C., Jonathan M. Eastman, Matthew W. Pennell, Pamela S. Soltis, Douglas E. Soltis, Cody E. Hinchliff, Joseph W. Brown, Emily B. Sessa, and Luke J. Harmon. "Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications." *New Phytologist* 207, no. 2 (2015): 454-467.

Taylor, Charlotte M., Roy E. Gereau, and Gretchen M. Walters. "Revision of Ancistrocladus Wall.(Ancistrocladaceae)." *Annals of the Missouri Botanical Garden* (2005): 360-399.

Tillyard, RJ. "A new classification of the order Perlaria." *The Canadian Entomologist* 53, no. 2 (1921): 35-43.

Theobald, Douglas L. "A formal test of the theory of universal common ancestry." *Nature* 465, no. 7295 (2010): 219.

Thomson, Ashley M., Oscar M. Vargas, and Christopher W. Dick. "Comparative analysis of 24 chloroplast genomes yields highly informative genetic markers for the Brazil nut family (Lecythidaceae)." *bioRxiv* (2017): 192112.

Thulin, Mats, Abigail J. Moore, Hesham El-Seedi, Anders Larsson, Pascal-Antoine Christin, and Erika J. Edwards. "Phylogeny and generic delimitation in Molluginaceae, new pigment data in Caryophyllales, and the new family Corbichoniaceae." *Taxon* 65, no. 4 (2016): 775-793.

Uribe-Convers, Simon, Justin R. Duke, Michael J. Moore, and David C. Tank. "A long PCR–based approach for DNA enrichment prior to next-generation sequencing for systematic studies." *Applications in plant sciences* 2, no. 1 (2014).

Van Dongen, Stijn Marinus. "Graph clustering by flow simulation." PhD diss., 2000.

Vargas, Oscar M., Edgardo M. Ortiz, and Beryl B. Simpson. "Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: Diplostephium)." *New Phytologist* 214, no. 4 (2017): 1736-1750.

Walker, Joseph F., Joseph W. Brown, and Stephen A. Smith. "Analyzing contentious relationships and outlier genes in phylogenomics." Systematic biology (2018).

Walker, Joseph F., Robert K. Jansen, Michael J. Zanis, and Nancy C. Emery. "Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes." *American journal of botany* (2015).

Walker, Joseph F., Ya Yang, Tao Feng, Alfonso Timoneda, Jessica Mikenas, Vera Hutchison, Caroline Edwards et al. "From cacti to carnivores: Improved phylotranscriptomic sampling and hierarchical homology inference provide further insight into the evolution of Caryophyllales." *American journal of botany* 105, no. 3 (2018): 446-462.

Walker, Joseph F., Ya Yang, Michael J. Moore, Jessica Mikenas, Alfonso Timoneda, Samuel F. Brockington, and Stephen A. Smith. "Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales." *American journal of botany* 104, no. 6 (2017): 858-867.

Walker, Joseph F., Michael J. Zanis, and Nancy C. Emery. "Comparative analysis of complete chloroplast genome sequence and inversion variation in Lasthenia burkei (Madieae, Asteraceae)." *American journal of botany* 101, no. 4 (2014): 722-729.

Wang, Ning, Ya Yang, Michael J. Moore, Samuel F. Brockington, Joseph F. Walker, Joseph W. Brown, Bin Liang et al. "Evolution of Portulacineae marked by gene tree conflict and gene family expansion associated with adaptation to harsh environments." *bioRxiv* (2018): 294546.

Washburn, Jacob D., James C. Schnable, Gavin C. Conant, Thomas P. Brutnell, Ying Shao, Yang Zhang, Martha Ludwig, Gerrit Davidse, and J. Chris Pires. "Genome-Guided Phylo-Transcriptomic Methods and the Nuclear Phylogentic Tree of the Paniceae Grasses." *Scientific Reports* 7, no. 1 (2017): 13528.

White, Philip J., Helen C. Bowen, Martin R. Broadley, Hamed A. El-Serehy, Konrad Neugebauer, Anna Taylor, Jacqueline A. Thompson, and Gladys Wright. "Evolutionary origins of abnormally large shoot sodium accumulation in nonsaline environments within the Caryophyllales." *New Phytologist* 214, no. 1 (2017): 284-293.

Wickett, Norman J., Siavash Mirarab, Nam Nguyen, Tandy Warnow, Eric Carpenter, Naim Matasci, Saravanaraj Ayyampalayam et al. "Phylotranscriptomic analysis of the origin and early diversification of land plants." *Proceedings of the National Academy of Sciences* 111, no. 45 (2014): E4859-E4868.

Wilson, Edward O. "A consistency test for phylogenies based on contemporaneous species." *Systematic zoology* 14, no. 3 (1965): 214-220.

Wolfe, Kenneth H., Wen-Hsiung Li, and Paul M. Sharp. "Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs." *Proceedings of the National Academy of Sciences* 84, no. 24 (1987): 9054-9058.

Xi, Zhenxiang, Liang Liu, Joshua S. Rest, and Charles C. Davis. "Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies." *Systematic biology* 63, no. 6 (2014): 919-932.

Yang, Ya, and Stephen A. Smith. "Optimizing de novo assembly of short-read RNA-seq data for phylogenomics." *BMC genomics* 14, no. 1 (2013): 328.

Yang, Ya, and Stephen A. Smith. "Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics." *Molecular Biology and Evolution* 31, no. 11 (2014): 3081-3092.

Yang, Ya, Michael J. Moore, Samuel F. Brockington, Jessica Mikenas, Julia Olivieri, Joseph F. Walker, and Stephen A. Smith. "Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events." *New Phytologist* 217, no. 2 (2018): 855-870.

Yang, Ya, Michael J. Moore, Samuel F. Brockington, Alfonso Timoneda, Tao Feng, Hannah E. Marx, Joseph F. Walker, and Stephen A. Smith. "An efficient field and laboratory workflow for plant phylotranscriptomic projects." *Applications in plant sciences* 5, no. 3 (2017).

Yang, Ya, Michael J. Moore, Samuel F. Brockington, Douglas E. Soltis, Gane Ka-Shu Wong, Eric J. Carpenter, Yong Zhang et al. "Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing." *Molecular Biology and Evolution* 32, no. 8 (2015): 2001-2014.

Yang, Ziheng. "PAML: a program package for phylogenetic analysis by maximum likelihood." *Bioinformatics* 13, no. 5 (1997): 555-556.

Yang, Ziheng. "On the best evolutionary rate for phylogenetic analysis." *Systematic Biology* 47, no. 1 (1998): 125-133.

Yang, Ziheng, Nick Goldman, and Adrian Friday. "Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem." *Systematic Biology* 44, no. 3 (1995): 384-399.

Yu, Yan, Qiuyun Xiang, Paul S. Manos, Douglas E. Soltis, Pamela S. Soltis, Bao-Hua Song, Shifeng Cheng, Xin Liu, and Gane Wong. "Whole-genome duplication and molecular evolution in Cornus L.(Cornaceae)–Insights from transcriptome sequences." *PloS one* 12, no. 2 (2017): e0171361.

Zakaria, Wan, Kok-Keong Loke, Muhammad-Mu'izzuddin Zulkapli, Mohd Salleh, Hoe-Han Goh, and Normah Mohd Noor. "RNA-seq analysis of Nepenthes ampullaria." *Frontiers in plant science* 6 (2016): 1229.

Zanis, Michael J., Douglas E. Soltis, Pamela S. Soltis, Sarah Mathews, and Michael J. Donoghue. "The root of the angiosperms revisited." *Proceedings of the National Academy of Sciences* 99, no. 10 (2002): 6848-6853.

Zuckerkandl, Emile, and Linus Pauling. "Molecules as documents of evolutionary history." *Journal of theoretical biology* 8, no. 2 (1965): 357-366.