Phylogenomic Perspectives on Evolutionary History: Examples from the Flowering Plant Lineage Ericales

by

Drew A. Larson

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
2022

Doctoral Committee:

Professor Christopher W. Dick, Co-chair
Associate Professor Stephen A. Smith, Co-chair
Associate Professor Daniel L. Rabosky
Associate Professor Selena Y. Smith

Drew A. Larson

larsonda@umich.edu

ORCID iD: 0000-0002-7557-9999

To my parents, for teaching me to work with purpose and explore with passion

Acknowledgements

I first thank **my parents, Mary and Larry Larson**, who have loved and supported me longer than anyone else, and who have always encouraged me to live well, study hard, and never stop learning about the world. Thanks also to **Amy Larson**, for always being a great sister and friend. I couldn't have asked a better family, or for a more wonderful childhood.

I thank the many teachers of the **Willmar Public School System**, who shaped my education during the first 18 years of my life and who helped prepare me for the opportunities ahead.

Thanks to all my professors at the **University of Minnesota, Morris**, especially:

Christopher T. Cole, whose brilliant courses in molecular biology, plant biology, and conservation biology expanded my understanding of the world and whose devotion to teaching inspires me to this day.

Peter H. Wyckoff, who first taught me what it meant to do research, believed in, and encouraged me as a scientist, and whose mentorship in many ways changed my life.

Margaret A. Kuchenreuther, who first sparked my interest in plant systematics, taught me the flora of my home state, and supervised my senior capstone presentation.

Thanks to both the **2016 EEB cohort** and my **field ecology classmates**, especially **Sasha Bishop, Zachary Hajian-Forooshani**, and **Peter Cerda**, for making the transition to graduate school very fun and making Ann Arbor feel like home.

I thank my fellow graduate students, **Jordan Bemmels**, **Caroline Parins-Fukuchi**, **Stephanie Alcala**, **Lijun Zhao**, **Hector Figueroa**, **William Weaver**, **Keyi Feng**, **Avinash Subramanian**, **Diana Medellín-Zabala**, **Janae White**, **Raquel Pizzardo**, **Lauren Hart**, and **Priscila Souza**, and everyone else who has been a part of the Smith and Dick Labs these last six years.

Thanks to **Deise Gonçalves**, **Jacob Berv**, **Oscar Vargas**, **Hannah Marx**, **Greg Stull**, **Ning Wang**, **Simon Uribe-Convers**, **Joseph Brown,** and the other postdocs that have taught me so much during graduate school.

I thank **Giorgia Auteri** and **Abby Kimmit** for their friendship, consistently good advice, and solid belays.

I thank **Nathanael Walker-Hale** for many laughs and his one-of-a-kind personality.

Thanks to **Raquel Rivadeneira** and **Jillian Myers** for their helpful advice in the lab on many occasions.

I thank **Brad Ruhfel** for teaching me so much about curating and managing collections during my time working at the University of Michigan Herbarium.

Thanks to the **EEB Graduate Student Coordinators**, **Cindy Carl**, **Kati Ellis**, and **Nathan Sadowsky** and all the **EEB Department staff** who helped make sure I met the formal requirements of the program.

I thank **Tim Utteridge**, **Jim Clarkson**, **Bill Baker**, **Alex Zuntini**, **Lorna Frankel**, **Lizzy Joice**, and everyone else who made me feel so welcome during my time working at Kew.

I thank **Beto Vicentini**, **Zé Luís Camargo**, **Paulo Assunção, Renata Pirani**, everyone else who helped make it possible for me to do fieldwork in Brazil.

I especially thank **Tamara Milton** for being a great friend and colleague since we first met during recruitment weekend. It has been wonderful to talk science, adventure, and spend time with you everywhere from Ann Arbor to the George Reserve, Miami, and the Amazon rainforest.

I especially thank **Joseph Walker** for being a wonderful mentor, co-author, and friend throughout my time in graduate school and for teaching me so much about phylogenetics and about life.

I give a very special thanks to **Caroline Edwards**, who has been there for me through it all, and who has shaped my work and life for the better in so many ways, both big and small.

I thank the members of my dissertation committee **Selena Smith** and **Dan Rabosky** for the insightful advice about my projects and career that they have shared with me over the years.

And finally, I thank my advisors **Chris Dick** and **Stephen Smith**. My perspectives on science have been profoundly shaped by them both. Throughout graduate school, they have encouraged me to work on what interested me and provided me with the opportunities, resources, and guidance that I needed to reach my scientific goals. They have both always treated me as a peer and valued my perspectives. This research would not have been possible without them and the lab environments they have fostered, and I cannot thank them enough for all they have done for me these last six years.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

This dissertation represents a series of advances in plant systematics and molecular phylogenetics. Processes such as whole genome duplication, historical introgression, and rapid diversification shape the genomes of plants. I investigate these processes in the flowering plant lineage Ericales, a morphologically disparate group that includes kiwifruits, pitcher plants, obligate parasites, and ecologically dominate tropical tree lineages.

Large genomic datasets offer the promise of resolving historically recalcitrant species relationships, but methods can yield conflicting results, especially when clades have experienced ancient, rapid diversification. In Chapter II, we analyzed the ancient radiation of Ericales and explored sources of uncertainty related to species tree inference, conflicting gene tree signal, and the inferred placement of gene and genome duplications. Support for relationships among major clades was inferred from multiple lines of evidence and was summarized in a consensus framework. Our results supported a history largely concordant with previous studies but suggests that paleopolyploidy may be responsible for the remaining uncertainty. Our broad sampling allowed us to place the position of a whole genome duplication before the radiation of most ericalean families.

Admixture is a mechanism by which populations of long-lived trees may acquire novel alleles. However, little is known about the genomes of most tropical tree species, or the extent to which they exchange genes. In Chapter III, we ask whether admixture occurs in an ecologically important clade of rainforest trees, the Parvifolia clade of *Eschweilera* (Lecythidaceae), which includes several of the most abundant tree species in Amazon forests. Using targeted sequence

capture for hundreds of individuals from across Lecythidaceae, we conducted a detailed phylogenomic investigation of the Parvifolia clade. We implement a novel workflow to test for admixture in target capture datasets. We found strong evidence of admixture among three ecologically dominant species but a lack of evidence for widespread genomic admixture in most lineages. Species were distinguishable from one another based on our sequencing targets, as was geographic structure within species.

Biogeography informs our understanding of patterns of global species diversity and the processes that shape them, but such inferences strongly rely on the quality of the genomic and fossil information employed. In Chapter IV, we use targeted sequence capture to collect data from species across Ericales, as well as the available fossil information, to investigate the origin and diversification of the primrose family (Primulaceae). We present updated phylogenetic and biogeographic hypotheses for the family and show that genomic evidence contradicts previous biogeographic inference based on morphology. Our results show that a major taxonomic revision of *Ardisia* and at least 19 closely related genera is required to circumscribe monophyletic genera.

While this dissertation advances our understanding of the evolution of Ericales, it has also revealed unanswered questions about the phylogeny of the order and the processes that have generated the groups diversity. Paleopolyploidy, rapid radiation, admixture, and long-distance dispersal are among the factors that have contributed to the evolution of the clade. Future work is needed to better characterize the relative importance of these factors and to continue refining the taxonomy of various clades within Ericales.

Chapter I

Introduction

A brief introduction to plant systematics

The field of plant systematics seeks to understand the diversity of plant life on Earth and develop systems of classifying this diversity in ways that are useful and informative. Many of the tenets of classification commonly used date back at least to the time of the Italian botanist Andreas Caesalpinus (1524–1603), who thought a system of classification should reflect "natural" groups, as well as aid in memory, and allow predictions about the properties of plants (Caesalpinus, 1583 as cited by Judd et al., 2015). Carl Linnaeus (1707–1778) developed the system of binomial nomenclature that continues to be used across scientific disciplines. The work of Antoine Laurent de Jussieu (1748–1836) established the basic system of hierarchical organization still used to classify plants by grouping genera into successively larger groups based on shared characteristics (Judd et al., 2015). For most of its history, plant systematics necessarily relied on readily observable morphological and anatomical traits as the basis for the description and differentiation of species and larger groups. Beliefs about how organisms should be grouped most "naturally" and differing ideas regarding which characteristics to prioritize in searching for breakpoints in the diversity of plant life led to large differences in the classification systems proposed by various authors (Judd et al., 2015).

The discovery that species are related to one another through descent with modification and therefore share genealogical history was perhaps most pivotal discovery in the history of biology. Charles Darwin and Alfred Russel Wallace independently came to recognize that biological evolution was the simplest explanation for the patterns of biodiversity that exist on Earth (Darwin, 1859; Darwin and Wallace, 1858). Evolution offered new light with which to classify life on Earth: groups of organisms could be defined as branches on the tree of life. A phylogeny, or evolutionary history, could form the basis of hierarchical classification, with organisms grouped together based on the degree of their shared ancestry.

Willi Hennig strongly argued that biological systematics should be phylogenetic. In his view, and the views of some of his contemporaries, systematists should seek to organize all living things into a system of hierarchical units based on evolutionary relationships, rather than based on overall similarity (Hennig, 1966, 1950). Therefore, named groups should be monophyletic groups or "clades", from the Greek word *Klados*, meaning branch. Monophyletic groups include a shared ancestral progenitor species and *all* lineages derived from that ancestor. Monophyly is now an indispensable criterion for most systematists.

It was Hennig who also first formalized ideas about which characteristics could be used to infer shared ancestry (Hennig, 1966, 1950). He showed that characters that are derived (apomorphies) relative to an ancestral condition and are shared by more than one species (synapomorphies) are the only phylogenetically informative traits that can be used to identify clades. In contrast, the overall number of similarities between organisms does not necessarily provide information on their degree of relatedness if the traits being considered are ancestral (plesiomorphic). During the same period of time, deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and the amino acids the comprise proteins had each been sequenced for the first time (reviewed by Stretton, 2002) and the idea to use molecular sequences for inferring evolutionary history was first published (Zuckerkandl and Pauling, 1965). Methods for identifying phylogenies that best explain the observed similarities among organisms in terms of parsimony and later maximum likelihood soon proliferated (Felsenstein, 1982, 1973).

The ability to read the sequences of biomolecules, and particularly nucleic acids, has revolutionized our understanding of biodiversity and led to the field of molecular systematics. This is because molecular sequences can provide unparalleled information about shared genetic characteristics that can be used to infer evolutionary relationships among organisms. Many early phylogenetic studies of plants focused on a small number of genes due to the expense of sequencing and computational limitations. Among the most used molecular sequences was the gene *rbcL*, which codes for the Ribulose-1,5-bisphosphate carboxylase-oxygenase protein in plants, critical for carbon fixation during photosynthesis. The internal transcribed spacer (ITS) region between ribosomal RNA genes has historically also been among the most important regions for plant molecular systematics, as well as for animal and fungal systematics. These and several other historically important DNA regions continue to provide useful information about phylogenetic relationships. More recently, advances in 1) sequencing technology, especially second generation sequencing, 2) ways of conducting reduced representation genomic sequencing (Cronn et al., 2012; Eaton and Ree, 2013; Johnson et al., 2018), 3) computational capacity, and 4) phylogenetic methods, have again revolutionized the information that can be brough to bear on questions about the evolutionary history of lineages.

By the 1990s, it had become clear based on molecular evidence that the historical systems of plant classification (e.g., Cronquist, 1981) had grouped plants together that were not monophyletic lineages. Excluding some descendants of a common ancestor renders a group paraphyletic, while grouping members of two distinct lineages results in a polyphyletic group. The first publication by the Angiosperm Phylogeny Group (APG I, 1998) sought to rectify the non-monophyly of broad groups by proposing new circumscriptions of many orders of angiosperms (flowering plants), so that all named orders were monophyletic to the extent possible based on existing data.

At present, organizing the diversity of plant life into monophyletic groups continues to be a major goal of plant taxonomists and other systematists. In classifying angiosperms, the hierarchical ranks of order, family, genus, and species are still commonly used in the naming of taxa, or groups of organisms. Some lineages are further classified into infraspecific taxa. A major criticism of classifying taxa in a ranked system is that it creates a false equivalency among groups with the same taxonomic rank. One might expect that botanical *families* have some defining feature that justifies their being assigned the rank of family, rather than that of genus or order. However, no such criteria are universally agreed upon and the rank assigned to clades is somewhat arbitrary (APG I, 1998). Thus, the taxonomic rank given to a clade is often the product of both historical context and nomenclatural convenience. A more recent, alternative system, PhyloCode, has also been proposed, which gives names to clades without ranks (Cantino and de Queiroz, 2020). However, the system of higher classification of plants outlined by APG (1998, 2003, 2009, 2016) continues to be the norm in the scientific literature.

Understanding the evolutionary history of life is foundational to our understanding of many aspects of the natural world. For example, phylogenies have become central to studies of biogeography, which asks where biodiversity exists and why. Accurate knowledge of phylogeny is also critical to our inferences of the evolution of novel morphological features, genome evolution, and polyploidy, and allows us to describe which lineages exchange genes with one another more accurately. Information about which lineages exist, where they live, and how they interact is also critical to efforts to conserve biological diversity, the accelerating loss of which has the potential to dramatically change the stability and functioning of the world's ecosystems. For these reasons and many others, systematics is, and will continue to be, central to our understanding of biology.

A taxonomic history of Ericales

The studies described in this dissertation focus on various aspects of the lineage of flowering plants called Ericales. This lineage offers many illustrative study systems within which to investigate patterns of plant biodiversity and the processes that govern their generation and maintenance. As circumscribed by APG IV (2016), Ericales is now considered to comprise 22 families, together containing more than 12,000 species (Stephens, 2001 onward). Ericales derives its name from the name given to the heath genus, *Erica* by Linneaus (1753), which was grouped in the family Ericae (which later became Ericaceae) by Jean Francois Durande in 1782 and by Antoine-Laurent de Jussieu in 1789 (International Plant Name Index, 2021). The named group that became the order Ericales is attributed to the Bohemian authors Friedrich von Berchtold and Jan Swatopluk Presl, who first referenced the group in their book "O přirozenosti rostlin" (On Plant Nature) which was published in Czech (Berchtold and Presl, 1820).

Members of Ericales were traditionally assigned to many distinct groups due to the dramatic diversity that has evolved since their common ancestor. Before molecular investigations became common, the widely used system by Cronquist (1981) placed its members into the separate orders Lecythidales, Ericales *sensu stricto* (*s.s.*), Diapensiales, Ebenales, and Primulales. In addition, the holoparasitic lineage Mitrastemonaceae was grouped with other parasitic plants in Rafflesiales, which we now know are only very distantly related (Judd et al., 2015). The impatiens family, Balsaminaceae was grouped with Geraniales, and the phlox family, Polemoniaceae, was included in Solanales. Early molecular studies by several authors clarified that this morphologically diverse group of plants were one another's closest extant relatives and re-circumscribed Ericales as a monophyletic group (APG I, 1998; Källersjö et al., 1998; Morton et al., 1996). While there is strong evidence that Ericales is now monophyletic as currently circumscribed, this is not necessarily true of other named groups within the order.

Concepts of species and genera and related terminology

The discussion of what constitutes a species is perennial among biologists (de Queiroz, 2007). A multitude of criteria exist for delimiting species boundaries, including biological species concepts (reproductive isolation), phylogenetic species concepts (various criteria involving monophyly), and evolutionary species concepts (independently evolving lineages). The delimitation and circumscription of species is not a main goal of this dissertation, though some of the results presented have implications for these boundaries.

Through careful study, systematists can identify lineages and the extent to which they are evolving independently, or non-independently, such as those that hybridize (Eaton et al., 2015; Eaton and Ree, 2013; Hardin, 1975) or obtain genes from other species through horizontal gene transfer (Davis and Xi, 2015). Species may also share genetic similarities due to incomplete sorting of ancestral variation during speciation, resulting in different parts of the genome having different evolutionary histories (Maddison, 1997). Phylogenetic network methods provide a framework for explicitly modeling the reticulate nature of genomes, and while not used in the studies presented here, are often invaluable for describing the evolutionary history of plants (Huson et al., 2005; Huson and Bryant, 2006).

Because plant lineages exhibit such a diverse spectrum of evolutionary processes, including hybridization, self-fertilization, agamospermy (viable seeds produced without fertilization), and polyploidy (whole genome duplication, often leading to reproductive isolation), there is no single definition of species that is useful to describe all plant lineages in all contexts (Judd et al., 2015). However, naming "species" provides a valuable, if provisional, means of describing what lineages we hypothesize to exist in nature. Here, the word species is generally used to refer to more or less independently evolving lineages, but that may still exchange genes with other species. In practice, this amounts to recognizing species boundaries as they are "generally accepted" by taxonomists and other plant systematists. This also generally corresponds to the information available in published monographs and databases, upon which the studies in this dissertation rely.

A somewhat different problem applies to the question of what should constitute a genus. Some early taxonomists, including Linnaeus, believed that genera were distinct, real entities that exist in nature, independent of any efforts to describe them (Judd et al., 2015). From an evolutionary perspective, however, there is no reason to decide that some lineages should be assigned the rank of genus, while others are not. Despite this, binomial taxon names are useful for communicating which lineage we are referring to, and for summarizing the biodiversity of life on Earth. There are many instances, for example, ecological investigations of tropical forests, where investigators may need to use units above the species level for individuals in their analyses, because confident species-level identification is not yet possible. Because they are so widely used, the genus names assigned to species should at least inform users as to what other lineages are most closely related. In this dissertation, I use only the criterion that genera and any other named group above the species level should be monophyletic according to the available data for a large majority of the relevant genomes. Rare hybridization among genera will still allow the resulting lineages to be assigned to the genus with which they share most of their genetic heritage. Conversely, frequent hybridization among lineages can be considered as evidence for recognizing those lineages as part of the same genus, since differentiating between the two is unlikely to be very useful. Based on this criterion, the co-authors of various chapters and I make several recommendations for genera that should be re-circumscribed in the future, to make all genera monophyletic.

Phylogenomics, modeling evolution, and a consensus approach

The word "phylogenomics" is now commonly used to refer to phylogenetic research that uses many gene regions comprising "genome-scale" data. There is no agreed upon definition of what "genome-scale" is, thus a phylogenomic dataset could be dozens of genes from chloroplast genomes, or hundreds to thousands of nuclear gene regions, or a combination of nuclear, plastid, and mitochondrial genes. Phylogenomic datasets can be used not only to investigate relationship among organisms, but also to ask about gene and genome duplication, rates of evolution, speciation, adaptation and selection, and the exchange of genetic material among lineages.

A major finding of phylogenomics has been the widespread occurrence of phylogenetic conflict among gene regions: phylogenetic histories that appear to be different in various molecular sequences from the same lineages (Walker et al., 2018b). The idea that phylogenetic signals can differ among genes has long been recognized, and can result from biological processes like hybridization, horizontal gene transfer, or incomplete lineage sorting (Maddison, 1997). Phylogenomic conflict can also result from systematic error, whereby issues such as violating the assumptions of evolutionary models, data quality, or misidentified orthology among sequences can cause misleading results (Walker et al., 2019). Differentiating between biological sources of phylogenetic conflict and systematic error is critical to our understanding of evolution. Indeed, cases where we can confidently infer that biological conflict exists can provide great insight into the processes that have generated the diversity of life on Earth (Green et al., 2010; Martin et al., 2006; Walker et al., 2018b, 2017). There are some cases where identifying systematic error is relatively straightforward, such as for plastid genomes that are not expected to experience much recombination (Doyle, 2022; Walker et al., 2019) while in other cases such as in large, polyploid, nuclear genomes, identifying systematic error is difficult.

Despite the size of modern phylogenomic datasets, individual gene regions can have large effects on the relationships inferred (Walker et al., 2018a). As shown in Chapters II-IV and elsewhere, the criteria used to process data, and the models used to analyze data can, impact the conclusions of phylogenomic analyses. All models have assumptions, many of which are routinely violated to some extent. For example, both hybridization and incomplete lineage sorting violate the assumptions of concatenated supermatrix analyses, which usually rely on the assumption that all the sequences in the matrix evolved along the same, bifurcating species tree. The substitution model may also rely on the assumption that all genes evolve at the same rate, related rates, or at independent rates. We know there is a possibility that some genes, or a subset of sequences within genes due to recombination, may have a phylogenetic history that truly differs from the species tree—we often even expect this to be the case (Maddison, 1997). However, concatenated supermatrix approaches are still commonly used, in part because they allow large amounts of data to be analyzed efficiently using relatively well-characterized evolutionary models and maximum likelihood statistical frameworks.

Alternatives to concatenation approaches also exist, including those that allow gene trees to differ from the species tree, such as is implemented in programs such as ASTRAL (Zhang et al., 2018). However, these approaches still rely on a variety of assumptions, including that the gene trees are accurate representations of all the sequence data used to generate the gene tree (i.e. that each "gene" has only one history), that each gene tree has a similar information content (or that you can accurately weight them), and that conflict among gene trees is due to incomplete lineage sorting and not hybridization or horizontal gene transfer. Inferring a gene tree requires assumptions that there is adequate information to estimate an appropriate substitution model or that you have chosen a substitution model that is adequate, that the sequences themselves are homologous, and that the individual nucleotides in each column of the alignment used to produce the gene trees are also homologous (i.e., are accurately aligned). Bayesian methods that use Markov chain Monte Carlo (MCMC) to estimate species trees also have many assumptions, though are not often discussed here because they are currently unable to accommodate very large genomic datasets.

Phylogenetic networks methods allow for reticulations in the evolutionary history of lineages, and so do not assume the species tree is strictly bifurcating (Huson and Bryant, 2006). However, to make these methods computationally tractable for estimating proportion of hybrid ancestry, most implementations require the user to specify the maximum number of reticulations allowed (Solís-Lemus et al., 2017; Than et al., 2008). These methods generally also assume that the gene trees or nucleotide matrices input are free of systematic errors and that reticulation events involved the specific lineages included in the analysis rather than unsampled or extinct lineages.

It is not always straightforward to determine which hypothesis is more strongly supported when multiple methods give distinct answers. In some cases, when the same dataset is used, information criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to select a preferred model, but even these require judgements about how heavily to penalize additional model parameters and can give different answers (e.g., Chapter IV). In other cases, such as when the phylogeny generated using maximum likelihood with a concatenated supermatrix conflicts with results from other tree methods, there is usually not a uniform, objective way to differentiate which hypothesis better reflects reality. When such cases occurred in the research for this dissertation, my co-authors and I have generally sought information from multiple approaches and looked for areas of consensus among different methodologies. When the inferred topology of the phylogeny is supported by multiple methods, we can perhaps have increased confidence that our hypothesis of relationships reflects an accurate evolutionary history of a majority fraction of the genomes we are studying. Similarly, in Chapter III, when investigating evidence for hybridization, we sought morphological evidence in the field, as well as multiple threads of genomic evidence for testing hypotheses of hybridization among species of *Eschweilera*. Disagreements among methods can conversely be taken as evidence of uncertainty, which might be especially important in phylogenomic studies, were the large amount of data included affects support metrics such that when model assumptions are violated one can find misleadingly high support for the wrong answer (Seo, 2008).

Historical context

I began my graduate studies in August of 2016, during a time of accelerating availability of genomic data for phylogenetics, which has continued to the present. The 1000 Plant

Transcriptomes Initiative (1KP) had made a large number of whole transcriptome sequences available (Matasci et al., 2014), but had not yet published their capstone paper (Leebens-Mack et al., 2019). A relatively new technology called target capture, targeted sequence capture, or target enrichment sequencing, which uses high-throughput “next generation sequencing” (sometime now called “second generation sequencing”) had been developed, but sequencing probes were only available for a handful of plant groups such as Asteraceae (Mandel et al., 2014) and Rosaceae (Liston, 2014). The human genome had first been sequenced over a decade ago (Venter et al., 2001), and more recent evidence had revealed that many people share genes with the extinct human lineage *Homo neandertalensis* (Green et al., 2010). New tools were also rapidly being developed to study the phylogenetic history of organisms at a broad scale.

At the outset, we knew which extant plant lineages formed the broader clade of Ericales, but there was still uncertainty in how the families were related to one another (APG IV, 2016). A series of papers had recently provided a first glimpse into the phylogenetic relationships among species of the Lecythidaceae based on Sanger sequencing of a few genes (Huang et al., 2015; Mori et al., 2015). Their results suggested that some species and genera were not monophyletic, including within the genus *Eschweilera*, which became the focus of Chapter III of this dissertation. Relationships within the family Primulaceae were known from studies using only a few genes. Notable among them were the results of Källersjö et al. (2000) who used sequences from three genes which together provided strong evidence that several traditionally recognized taxa were non-monophyletic and formed the basis for subsequent taxonomic revisions of the family. The genus *Maesa*, which had usually been regarded as the sole member of a subfamily of Myrsinaceae for over 100 years, was moved to its own family because it was clear it did not form a monophyletic group with other members of Myrsinaceae (Anderberg et al., 2000). Subsequent revisions to Myrsinaceae had followed as well (Manns and Anderberg, 2009).

Data collection for Chapter II of this dissertation began in earnest in January of 2018. A preprint version of that manuscript was first published in November of 2019. A later version of Chapter II was published in *American Journal of Botany* in April of 2020. We began making field collections of Lecythidaceae in Brazil for Chapter III in February of 2018 and submitted that manuscript for publication in April of 2021. Chapter III was first published by *New Phytologist* in August of 2021. I began sampling Primulaceae and other Ericales at the Kew Herbarium and preparing target capture sequencing libraries as part of the Plant and Fungal Tree of Life Project (PAFTOL) in August of 2019. I also sampled additional Primulaceae at the University of Michigan Herbarium in January of 2021. The results of this work on Primulaceae appear for the first time in Chapter IV of this dissertation.

Goals of this dissertation and chapter summaries

The goal of this dissertation is to advance our understanding of plant systematics and molecular phylogenetics through studies on various branches of the Ericales phylogeny. Chapters II-IV are each informed by the methods of molecular phylogenetics, the application of novel bioinformatic workflows, and the rich base of botanical knowledge that has been developed by previous investigators. It is my hope that the results of these studies will further our understanding of how evolution has shaped the diversity of plant forms that now exist.

Chapter II is an investigation of the relationships among the families of Ericales using sequences from hundreds of coding genes derived from transcriptomes. In this chapter, we also investigate the relative support for several possible relationships and discuss the limits of what we can infer about ancient divergence events in the presence of large numbers of gene duplications, even using large genomic datasets. Chapter III focuses on the Brazil nut family, Lecythidaceae, and seeks to shed light on an important and largely unanswered question about tropical biodiversity: do closely related, coexisting tree species hybridize with one another? We found evidence that some of the most abundant tree species in the Amazon do indeed hybridize but are still genetically distinguishable from one another. Chapter IV is the first phylogenomic investigation focused on the primrose family (Primulaceae). Previous phylogenetic studies of Primulaceae had suggested several taxa that might be non-monophyletic, but lack of phylogenetic resolution hindered our understanding of the group's evolution. We show that by using a phylogenomic approach, we can better understand the biogeographic history of the clade and confidently show that many revisions to the taxonomy of the group are necessary to make all genera reflective of the group's evolutionary history.

Chapter II

A Consensus Phylogenomic Approach Highlights Paleopolyploid and Rapid Radiation in the History of Ericales

Preamble: This chapter has been published in the *American Journal of Botany*. The citation for this chapter is: Larson, D.A., Walker, J.F., Vargas, O.M., Smith, S.A., 2020. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. American Journal of Botany 107, 773–789. https://doi.org/10.1002/ajb2.1469.

Abstract

Large genomic data sets offer the promise of resolving historically recalcitrant species relationships. However, different methodologies can yield conflicting results, especially when clades have experienced ancient, rapid diversification. Here, we analyzed the ancient radiation of Ericales and explored sources of uncertainty related to species tree inference, conflicting gene tree signal, and the inferred placement of gene and genome duplications. We used a hierarchical clustering approach, with tree-based homology and orthology detection, to generate six filtered phylogenomic matrices consisting of data from 97 transcriptomes and genomes. Support for species relationships was inferred from multiple lines of evidence including shared gene duplications, gene tree conflict, gene-wise edge-based analyses, concatenation, and coalescent-based methods, and is summarized in a consensus framework.

Our consensus approach supported a topology largely concordant with previous studies, but suggests that the data are not capable of resolving several ancient relationships because of lack of informative characters, sensitivity to methodology, and extensive gene tree conflict correlated with paleopolyploidy. We found evidence of a whole-genome duplication before the radiation of all or most ericalean families, and demonstrate that tree topology and heterogeneous evolutionary rates affect the inferred placement of genome duplications. We provide several hypotheses regarding the history of Ericales, and confidently resolve most nodes, but demonstrate that a series of ancient divergences are unresolvable with these data. Whether paleopolyploidy is a major source of the observed phylogenetic conflict warrants further investigation.

Introduction

The flowering plant clade Ericales contains several ecologically important lineages that shape the structure and function of ecosystems including tropical rainforests (e.g., Lecythidaceae, Sapotaceae, Ebenaceae), heathlands (e.g., Ericaceae), and open habitats (e.g., Primulaceae, Polemoniaceae) around the globe (He et al., 2014; Hedwall et al., 2013; Memiaghe et al., 2016; Moquet et al., 2017; ter Steege et al., 2006). With 22 families comprising ca. 12,000 species (APG IV, 2016 ; Stevens, 2001 onward), Ericales is a diverse and disparate clade with an array of economically and culturally important plants. These include agricultural crops such as blueberries (Ericaceae), kiwifruits (Actinidiaceae), sapotas (Sapotaceae), Brazil nuts (Lecythidaceae), and tea (Theaceae), as well as ornamental plants such as cyclamens and primroses (Primulaceae), rhododendrons (Ericaceae), and phloxes (Polemoniaceae). Holoparasitism has arisen at least twice in Ericales (Monotropoideae:Ericaceae, Mitrastemonaceae), as has carnivory in the American pitcher plants (Sarraceniaceae). Although the florally disparate Ericales has been a well-recognized clade since the widespread implementation of phylogenetic methods (Anderberg et al., 2002; Chase et al., 1993; Leebens-Mack et al., 2019; Rose et al., 2018; Schönenberger et al., 2005), the evolutionary relationships among major clades within Ericales remain contentious.

One of the first molecular studies investigating these deep relationships used three plastid and two mitochondrial loci; the authors concluded that the data set was unable to resolve several interfamilial relationships (Anderberg et al., 2002). These relationships were revisited by Schönenberger et al. (2005) with 11 loci (two nuclear, two mitochondrial, and seven chloroplast). Schönenberger et al. (2005) found that maximum parsimony and Bayesian analyses provided support for the resolution of some early diverging lineages. Rose et al. (2018) utilized three nuclear, nine mitochondrial, and 13 chloroplast loci in a concatenated supermatrix consisting of 49,435 aligned sites and including 4531 ericalean species but with 87.6% missing data. Despite the extensive taxon sampling utilized in Rose et al. (2018), several relationships were only poorly supported, including several deep divergences that the authors show to be the result of an ancient, rapid radiation. The 1KP initiative (Leebens-Mack et al., 2019) analyzed transcriptomes from across green plants, including 25 species of Ericales; their results suggest that whole-genome duplications (WGDs) have occurred several times in ericalean taxa. However, the equivocal support they recover for several interfamilial relationships within Ericales highlights the need for a more thorough investigation of the biological and methodological sources of phylogenetic incongruence in the group (Anderberg et al., 2002; Bremer et al., 2002; Leebens-Mack et al., 2019; Rose et al., 2018; Schönenberger et al., 2005).

Despite the increasing availability of genome-scale data sets, many relationships across the Tree of Life remain controversial. Research groups recover different answers to the same evolutionary questions, often with seemingly strong support (e.g., Shen et al., 2017). One benefit of genome-scale data for phylogenetics (i.e., phylogenomics) is the ability to examine conflicting signals within and among data sets, which can be used to help understand conflicting species tree results and conduct increasingly comprehensive investigations as to why some relationships remain elusive. A key finding in the phylogenomics literature has been the high prevalence of conflicting phylogenetic signals among genes at contentious nodes (Brown et al., 2017b; Reddy et al., 2017; Vargas et al., 2017; Walker et al., 2017). Such conflict may be the result of biological processes (e.g., introgression, incomplete lineage sorting, horizontal gene transfer), but can also occur because of lack of phylogenetic information or other methodological artifacts (Eaton et al., 2015; Richards et al., 2018; Walker et al., 2019). By identifying regions of high conflict, it becomes possible to determine areas of the phylogeny where additional analyses are warranted and future sampling efforts might prove useful. Transcriptomes provide information from hundreds to thousands of coding sequences per sample and have elucidated many of the most contentious relationships in the green plant phylogeny (e.g., Leebens-Mack et al., 2019; Simon et al., 2012; Walker et al., 2017; Wickett et al., 2014). Transcriptomes also provide information about gene and genome duplications not provided by most other common sequencing protocols for nonmodel organisms. Gene duplications are often associated with important molecular evolutionary events in taxa, but duplicated genes are also inherited through descent and should therefore contain evidence for how clades are related. By leveraging the multiple lines of phylogenetic evidence and the large amount of data available from transcriptomes, several phylogenetic hypotheses can be generated and tested to gain a holistic understanding of contentious nodes in the Tree of Life.

In this study, we sought to understand the evolutionary history of Ericales by analyzing sequences from thousands of homolog clusters to investigate support for contentious interfamilial relationships, ancient gene and genome duplications, heterogeneous rates of evolution, and conflicting signals among genes. We examined the deep relationships of Ericales to determine whether the data strongly support any resolution. We also consider the possibility that the data are unable to resolve these relationships (e.g., a hard or soft polytomy) despite previous strong support for alternative resolutions and thousands of transcriptomic sequences. While future developments in methods and sampling will likely continue to elucidate many contentious relationships across the plant phylogeny, we consider whether a polytomy may represent a more justifiable representation of the evolutionary history of a clade than any single fully bifurcating species tree given our current resources.

In applying a phylogenetic consensus approach that considers several methodological alternatives, we examined disagreement among methods. We explore this approach as a means of providing valuable information about whether the available data may be insufficient to confidently resolve a single, bifurcating species tree, even if a given methodology may suggest a resolved topology with strong support. Our investigation of the evolution of Ericales may be particularly well-suited to initiate a discussion about the affect of topological uncertainty on our ability to confidently resolve the placement of rare evolutionary events (e.g., whole-genome duplication, major morphological innovation), the prevalence of biological polytomies across the Tree of Life, and when polytomies may be considered useful representations of evolutionary relationships in the postgenomic era.

Materials and Methods

Taxon sampling and de novo assembly procedures

We assembled a data set consisting of coding sequence data from 97 transcriptomes and genomes—the most extensive exome data set for Ericales to date (Appendix A). The data set was constructed by obtaining the ericalean transcriptomes included in Matasci et al. (2014) and Vargas et al. (2019). In addition, at least one sequencing run for every species with data available on the Sequence Read Archive was downloaded, with preference given to those with the most reads (Leinonen et al., 2010). Outgroup sampling included cornalean transcriptomes from Matasci et al. (2014) and several reference genomes from Ensembl Plants (http://lants.ensembl.org). Samples assembled from raw reads were done so with Trinity version 2.5.1 using default parameters and the "–trimmomatic" option (Grabherr et al., 2011; Leinonen et al., 2010; Matasci et al., 2014). Assembled reads were translated using TransDecoder v5.0.2 (Haas et al., 2013) with the option to retain Basic Local Alignment Search Tool (BLAST) hits to a custom protein database consisting of *Beta vulgaris* (Amaranthaceae), *Arabidopsis thaliana* (Brassicaceae), and *Daucus carota* (Apiaceae) obtained from Ensembl plants (Altschul et al., 1990). The translated amino acid sequences from each assembly were reduced by clustering with CD-HIT version 4.6 with settings -c 0.995 -n 5 (Fu et al., 2012). As a quality control step, nucleotide sequences for the chloroplast genes *rbcL* and *matK* were extracted from each assembly using a custom script (see Data Availability Statement) and then queried using BLAST against the National Center of Biotechnology Information (NCBI) online database (Altschul et al., 1990). In cases where the top hits were to a species not closely related to that of the query, additional sequences were investigated and if contamination, misidentification, or other issues seemed likely, the transcriptome was not included in further analyses. The final transcriptome sampling included 86 ingroup taxa spanning 17 of the 22 families within Ericales as recognized by the Angiosperm Phylogeny Group (Appendix A; APG IV, 2016).

Homology inference

We used the hierarchical clustering procedure from Walker et al. (2018b) for homology identification. In short, the method involves performing an all-by-all BLAST procedure on user-defined clades; homolog clusters identified within each clade are then combined recursively with clusters of a sister clade based on sequence similarity until clusters from all clades have been combined. To assign taxa to groups for clustering, we identified coding sequences from the genes *rpoC2*, *rbcL*, *nhdF*, and *matK* using the same script as for the quality control step. Sequences from each gene were then aligned with MAFFT v7.271 (Katoh et al., 2002; Katoh and Standley, 2013). The alignments were concatenated using the command pxcat in phyx (Brown et al., 2017a) and the resulting supermatrix was used to estimate a species tree with RAxML v8.2.12 (Stamatakis, 2014). The inferred tree was used to manually assign taxa to one of eight clades for initial homolog clustering (Appendix A). Homolog clustering within each group was performed following the methods of Yang and Smith (2014).

Nucleotide sequences for the inferred homologs within each group were aligned with MAFFT v7.271 and columns with less than 10% occupancy were removed with the pxclsq command in phyx. Homolog trees were estimated with RAxML, unless the homolog had >500 tips, in which case FastTree v2.1.8 was used (Price et al., 2010). Tips with branch lengths longer than 1.5 substitutions per site were trimmed because the presence of highly divergent sequences in homolog clusters is often the result of misidentified homology (Yang and Smith, 2014). When a clade is formed of sequences from a single taxon, it likely represents either in-paralogs or alternative splice sites and because neither of these provide phylogenetic information, we retained only the tip with the longest sequence, excluding gaps introduced by alignment. This procedure was repeated twice with refined clusters using the same settings. Homologs among each of the eight groups were then recursively combined (https://github.com/jfwalker/Clustering) and homolog trees were again estimated and refined with the same settings except that internal branches longer than 1.5 substitutions per site were also cut—again to reduce potentially misidentified homology. Homolog trees were again re-estimated and refined by cutting terminal branches longer than 0.8 substitutions per site and internal branches longer than 1.0 substitutions per site. The final homolog set contained 9469 clusters.

Initial ortholog identification and species tree estimation

Orthologs were extracted from homolog trees using the rooted tree (RT) method, which allows for robust orthology detection even after genome duplications (Yang and Smith, 2014). Because our interest is in addressing phylogenetic questions, rather than those of gene functionality, here we use the term *ortholog* to describe clusters of sequences that have been inferred to be monophyletic based on their position within inferred homolog trees after accounting for gene duplication. For the RT procedure, seven cornalean taxa as well as *Arabidopsis thaliana* and *Beta vulgaris* were used as outgroups (Appendix A). Previous work has suggested that Ericales is sister to the euasterids with Cornales sister to Ericales + euasterids (Stevens, 2001 onward), therefore we treated *Helianthus annuus* (Asteraceae) and *Solanum lycopersicum* (Solanaceae) as ingroup taxa for the purposes of the RT procedure so that orthologs could be rooted on a non-ericalean taxon after ortholog identification. Orthologs were not extracted from homolog groups with more than 5000 tips because of the uncertainty in reconstructing very large homolog trees (Walker et al., 2018b). Orthologs with sequences from fewer than 50 ingroup taxa were discarded to reduce the amount of missing data in downstream analyses. The tree-aware ortholog identification employed here should provide the best available safeguard against misidentified orthology, which could mislead phylogenetic analyses (Brown and Thomson, 2017; Eisen, 1998; Gabaldón, 2008; Yang and Smith, 2014).

The resulting ortholog trees were then filtered to require at least one euasterid taxon, *Helianthus annuus* or *Solanum lycopersicum*, for use as an outgroup for rooting within each ortholog. If both outgroups were present in an ortholog tree but no bipartition existed with only those taxa (i.e., the tree could not be rooted on both), the ortholog was discarded because the monophyly of the euasterids is well established. Terminal branches longer than 0.8 were again trimmed, resulting in a refined data set containing 387 orthologs. Final nucleotide alignments were estimated with PRANK v.150803 (Löytynoja, 2014) and cleaned for a minimum of 30% column occupancy using the pxclsq function in phyx. Alignments for the 387 orthologs were concatenated using pxcat in phyx and a maximum likelihood (ML) species tree and rapid bootstrap support (200 replicates) was inferred using RAxML v8.2.4 with the command raxmlHPC-PTHREADS, the option -f a, and a separate GTRCAT model of evolution estimated for each ortholog; this resulted in a topology we refer to as the maximum likelihood topology (MLT).

To more fully characterize the likelihood space for these data, 200 regular (i.e., nonrapid) bootstraps with and without the MLT as a starting tree were conducted in RAxML using the -b option in separate runs. To investigate the possible effect of ML tree search algorithm on phylogenetic inference for the 387 ortholog supermatrix, two additional ML trees were estimated, one using RAxML v8.2.11 and the command raxmlHPC-PTHREADS-AVX and the other with IQ-TREE v1.6.1 and the options -m GTR+Γ -n 0 and model partitions for each ortholog specified with the -q option (Nguyen et al., 2015). Likelihood scores of the best scoring tree for each of these tree search algorithms were compared to determine whether the MLT was the topology with the best likelihood score. Trees for each individual ortholog were estimated individually using RAxML, with the GTRCAT model of evolution and 200 rapid bootstraps using the option -f a. A coalescent-based maximum quartet support species tree (MQSST) was estimated using ASTRAL v5.6.2 (Zhang et al., 2018) with the resulting 387 ortholog trees.

Gene-wise log-likelihood comparisons

A comparison of gene-wise likelihood support for the MLT against a conflicting backbone topology recovered in all 200 rapid bootstrap replicates (i.e., the rapid bootstrap topology, RBT) was conducted using a two-topology analysis (Shen et al., 2017; Walker et al., 2018a). We chose to investigate the RBT topology because even though the MLT received the best likelihood score recovered by any of the tree search algorithms, the MLT was never recovered by RAxML rapid bootstrap replicates, suggesting that the MLT was not broadly supported in likelihood space. A three-topology comparison was also conducted by calculating the gene-wise log-likelihoods of a third topology where the MLT was modified such that the interfamilial backbone relationships were those recovered in the ASTRAL topology (AT). This constructed tree was used in lieu of the actual ASTRAL topology to minimize the effect of intrafamilial topological differences on likelihood calculations. In both tests, the ML score for each gene is calculated while constraining the topology under multiple alternatives. Branch lengths were optimized for each input topology and GTR+Γ was optimized for each supermatrix partition (i.e., each ortholog) for each topology (Shen et al., 2017; Walker et al., 2018a). Results from both comparisons were visualized using a custom R script (R Core Team, 2019).

An edge-based analysis was performed to compare the likelihoods of competing topologies while allowing gene tree conflict to exist outside the relationship of interest (Walker et al., 2018a). Our protocol was similar to that of Walker et al. (2018a), except that instead of a defined "TREE SET" we used constraint trees with clades defining key bipartitions corresponding to each of three competing topologies using the program EdgeTest.py (Walker et al., 2019). The likelihood for each gene was calculated using RAxML-ng with the GTR+Γ model of evolution, using the brlopt nr_safe option and an epsilon value of 1×10^{-6} while a given relationship was constrained (Kozlov et al., 2019). The log-likelihood of each gene was then summed to give a likelihood score for that relationship. Because of the equivocal position of Ebenaceae recovered by bootstrapping, an additional edge-based analysis was conducted to investigate the placement of clade using the program Phyckle (Smith et al., 2020), which uses a supermatrix and a set of constraint trees specifying conflicting relationships and reports, for each gene, the ML as well as the difference between the best and second-best topology. This allows quantification of gene-wise support at a single edge as well as how strongly the relationship is supported in terms of likelihood.

Gene duplication comparative analysis

Homolog clusters were filtered such that sequences shorter than half the median length of their cluster and clusters with more than 4000 sequences were removed to minimize artifacts due to uncertainty in homolog tree estimation. Homolog trees were then re-estimated with IQ-TREE with the GTR+Γ model and SH-aLRT support followed by cutting of internal branches longer than 1.0 and terminal branches longer than 0.8 inferred substitutions per site. Rooted ingroup clades were extracted with the procedure from Yang et al. (2015) with all non-ericalean taxa as outgroups. Gene duplications were inferred with the program phyparts, requiring at least 50% SH-aLRT support to avoid the inclusion of very poorly supported would-be duplications (Smith et al., 2015). By mapping gene duplications in this way to competing topological hypotheses (i.e., the MLT, RBT, and AT), as well as several hypothetical topologies employed to reveal the number of gene duplications uniquely shared between clades that were never recovered as sister in the species trees, we determined the number of duplications uniquely shared among several ericalean clades.

When using tree-based methods to infer the placement of gene duplications, the inferred location of duplications depends on the species tree topology. Therefore, a gene duplication can map to a different edge in the species tree than expected based on the homolog tree because of conflict between the homolog and species trees. For example, if in a homolog tree, taxon A and taxon B share a gene duplication, but in the species tree taxon C is sister to taxon A, and taxon B sister to those, then the duplication will be mapped to the branch that includes all three taxa, because the duplication is mapped to the smallest clade that includes both taxon A and taxon B. However, if the duplication is instead mapped to a species tree where A and B are sister, the relevant duplication would instead be mapped to the correct, more exclusive branch that specifies the clade for which there is evidence of that duplication in the homolog tree (i.e., only taxa A and B). Once it has been determined how many duplications are actually supported by the homolog trees, comparisons between competing species relationships can be made. It is important to note that incomplete taxon sampling and other biases should be considered when applying such a comparative test of gene duplication number. Assuming there are duplicated genes present in the taxa of interest, clades with a greater number of sampled taxa or more complete transcriptomes will likely share more duplications simply because of the fact that more genes will have been sampled in the data set. Therefore, we investigate the use of gene duplications shared among clades as an additional, relative metric of topological support capable of corroborating other results (i.e., that if clades share many gene duplications unique to them, they are more likely to be closely related), while recognizing that the absolute number of duplications shared by various clades are affected by imperfect sampling.

Synthesizing support for competing topologies

We reviewed support for each of the three main topological hypotheses (i.e., MLT, RBT, and AT) and determined the most commonly supported interfamilial backbone. Because the comparative gene duplication analysis and constraint tree analyses both supported the RBT over other candidates, and the MLT was found to occupy a narrow peak in likelihood space based on bootstrapping and a two-topology test, the RBT was the most commonly supported backbone and was further explored with additional measures of support. Quartet support was assessed on the RBT using the program Quartet Sampling with 1000 replicates (Pease et al., 2018). We used this procedure to measure quartet concordance (QC), quartet differential (QD), quartet informativeness (QI), and quartet fidelity (QF). Briefly, QC measures how often the concordant relationship is recovered with respect to other possible relationships, QD helps identify if a relationship has a dominant alternative, and QI corresponds to the ability of the data to resolve a relationship of interest, where a quartet that is at least 2.0 log-likelihood (LL) better than the alternatives is considered informative. Finally, QF aids in identifying rogue taxa (Pease et al., 2018). Gene conflict with the corroborated topology was assessed using the bipartition method as implemented in phyparts using gene trees for the 387 orthologs, which were first rooted on the outgroups *Helianthus annuus* and *Solanum lycopersicum* using a ranked approach with the phyx command pxrr. Gene conflict was assessed both requiring node support (BS ≥ 70) and without any support requirement; the support requirement should help reduce noise in the analysis, but we also ran the analysis without a support requirement to ensure that potentially credible (but only partially supported) bipartitions were not overlooked. Results for both were visualized using the program phypartspiecharts (https://github.com/mossmatters/phyloscripts/).

Expanded ortholog data sets

In order to explore the impact of data set construction, orthologs were inferred from the homolog trees as for the 387 ortholog set but with modifications to taxon requirements and refinement procedures described below. For each data set, sequences were aligned separately with both MAFFT and PRANK and cleaned for 30% occupancy. A supermatrix was constructed with each and an ML tree was estimated with IQ-TREE with and without a separate GTR+Γ model partition for each ortholog to test the effect of model on phylogenetic inference. Individual ortholog trees were estimated with RAxML and used to construct an MQSST with ASTRAL.

2045 ortholog set—Orthologs were filtered such that there was no minimum number of taxa and at least two tips from each of the following five groups: (1) Primulaceae; (2) Polemoniaceae and Fouquieriaceae; (3) Lecythidaceae; (4) outgroups including *Solanum* and *Helianthus* as well as Marcgraviaceae and Balsaminaceae (i.e., the earliest diverging ericalean clade in all previous analyses); and (5) all other taxa. This filtering resulted in a data set with 2045 orthologs.

Ortholog tree support for conflicting placements of Ebenaceae was assessed for the PRANK-aligned orthologs using Phyckle.

4682 ortholog set—Orthologs were not filtered for any taxon requirements. Sequences were aligned with MAFFT, ortholog trees were estimated with RAxML, and terminal branches longer than 0.8 were trimmed. Sequences were then realigned separately with MAFFT and PRANK, and cleaned as before.

1899, 661, and 449 ortholog sets—To assess the effect of requiring *Helianthus* and *Solanum* outgroups in the 387 ortholog set and to further explore the effect of taxon requirements on the inferred topology, each homolog tree that produced an ortholog in that data set was re-estimated in IQ-TREE with SH-aLRT support. Calculating this support allowed visual assessment of whether uncertain homolog tree construction was affecting the ortholog identification process. This did not appear to be a major issue because orthologs (i.e., clades of ingroup tips subtended by outgroup tips) within homolog trees typically received strong support as monophyletic. Following homolog tree re-estimation, orthologs were identified as described above with no minimum taxa requirement. Sequences were aligned with MAFFT and any taxon with >75% missing data for a given ortholog was removed. Filtered alignments were then realigned separately with both MAFFT and PRANK, and cleaned for 30% column occupancy. This resulted in a data set with 1899 orthologs. To investigate the influence of taxon requirements, two subsets of this 1899 ortholog set were generated by requiring a minimum of 30 and 50 taxa, resulting in 661 and 449 orthologs, respectively. Gene tree conflict was assessed for the 449 MAFFT-aligned data set by rooting all ortholog trees on all taxa in Balsaminaceae and

Marcgraviaceae (i.e., the balsaminoid Ericales) because all previous analyses showed this clade to be sister to the rest of Ericales; phyparts was then used to map ortholog tree bipartitions to the ML tree for this data set and the results were visualized with phypartspiecharts.

Estimating substitutions supporting contentious clades

To estimate the signal present in ortholog alignments informing various relationships, we developed a procedure that identifies clades of interest within an ortholog tree and uses the estimated branch length leading to that clade, multiplied by the length of the corresponding sequence alignment, to estimate the number of substitutions implied by that branch. Applying this approach to the trees from the 449 ortholog MAFFT-aligned data set with appropriate taxon sampling to allow rooting on a member of the balsaminoid Ericales, we calculated the approximate number of substitutions that are inferred to have occurred along the branch leading to the most recent common ancestor (MRCA) of two or more of the following clades: Primulaceae, Polemoniaceae + Fouquieriaceae, Lecythidaceae, Ebenaceae, and Sapotaceae. In addition, we assessed substitution support for several noncontroversial relationships, namely that each of the following was monophyletic: Polemoniaceae + Fouquieriaceae, Lecythidaceae, and the non-balsaminoid Ericales. Mean and median values for substitution support were calculated and a distribution of these values was plotted using custom R scripts.

Synthesis of uncertainty, consensus topology, and genome duplication inference

We took into consideration all previous results, including those of the expanded data sets, gene tree conflict, and substitution support, to determine which relationships were generally well-supported by the results of this study, and which were not. In cases where an interfamilial or intrafamilial relationship remained irresolvable when considering the preponderance of the evidence (i.e., was not supported by a majority of methods employed after accounting for nested, conflicting relationships), that relationship was not included in the consensus topology. Gene duplications were mapped to the consensus topology using the methods of the comparative duplication analysis described above. Internodes on inferred species trees with notably high numbers of gene duplications were used as one line of evidence for assessing putative WGDs. Further investigation into possible genome duplications was conducted by plotting the number of synonymous substitutions (K_s) between paralogs according previously published methods (Yang et al., 2015) after removing sequences shorter than half the median length of their cluster. Multispecies K_s values (i.e., ortholog divergence K_s values) for selected combinations of taxa were generated using previously published methods (Wang et al., 2019). The effect of evolutionary rate-heterogeneity among ericalean species was investigated by conducting a multispecies $K_s$analysis of each non-balsaminoid ericalean taxon against *Impatiens balsamifera*, because all evidence suggests each of these species pairs have the same MCRA (i.e., the deepest node in the Ericales phylogeny). Because each pair has the same MRCA, the resulting ortholog peak in each case, represents the same speciation event, and differences in the location of this peak among K_s plots are the result of differences in evolutionary rate among species. In rare cases where the location of the ortholog peak was ambiguous (i.e., these were two or more local maxima near the global maximum) the plot was not considered in the rate-heterogeneity analysis. All single and multispecies K_s plots were generated using custom R scripts.

Results

Initial 387 ortholog data set

The concatenated supermatrix for the 387 ortholog data set contained 441,819 aligned sites with 76.1% ortholog occupancy and 57.8% matrix occupancy. The MLT recovered by RAxML v8.2.4 is shown in Figure 2-1. The 200 rapid bootstrap trees all contained the same backbone topology (i.e., the RBT) that differed from that of the MLT by one relationship. However, regardless of which tree search algorithm was used, the likelihood score for the MLT was better than for any other topology recovered for this data set. This indicates that while the MLT is only supported by a narrow peak in likelihood space, it was indeed the topology with the best likelihood based on the methods employed for the 387 ortholog data set. The MLT contained the clade Polemoniaceae + Fouquieriaceae as sister to a clade consisting of Ebenaceae, Sapotaceae, and what is referred to here as the "Core" Ericales: the clade that includes Actinidiaceae, Diapensiaceae, Ericaceae, Pentaphylacaceae, Roridulaceae, Sarraceniaceae, Styracaceae, Symplocaceae, and Theaceae. The RBT instead placed Lecythidaceae sister to Ebenaceae, Sapotaceae, and the Core (a hypothetical clade herein referred to as ESC), which was recovered by 100% of rapid bootstrap replicates (Figure 2-1). A gene-wise log-likelihood analysis comparing these two topologies is shown in Appendix A. The cumulative log-likelihood difference between the MLT and RBT was approximately 3.57 in favor of the MLT; there were 27 orthologs that supported the MLT over the RBT by a score larger than this, and the exclusion of any of these from the supermatrix could cause the RBT to become the topology with the best likelihood. Both of these topologies were considered as candidates in a search for a corroborated interfamilial backbone. Regular (i.e., non-rapid) bootstrapping in RAxML resulted in 6.5% support for Polemoniaceae + Fouquieriaceae sister to ESC (i.e., the MLT) unless the MLT was given as a starting tree, under which conditions the MLT topology received 100% bootstrap support. Unguided, regular bootstrapping instead suggested strong support (90%) for a clade consisting of Lecythidaceae and ESC (Appendix A). Interfamilial relationships recovered in the ASTRAL topology (AT) were congruent with those of the MLT except that Primulaceae was recovered as sister to Lecythidaceae + ESC, but with only 54% local posterior probability (Appendix A). A total of seven intrafamilial relationships differed between the AT and MLT. The AT was considered as a third candidate for an interfamilial backbone.

There was an effect of the ML tree search algorithm that may be important to note for future phylogenomic studies (Zhou et al., 2017). Reconducting an ML tree search for the 387 ortholog supermatrix with RAxML v8.2.11 using HPC-PTHREADS-AVX architecture and -m GTRCAT or with IQ-TREE and GTR+Γ resulted in a species tree with a different topology than under any other conditions for the 387 ortholog data set. To ensure direct comparability of scores, we recalculated the likelihood of both RAxML trees with IQ-TREE using the options show-lh, -te, and -blfix. The log-likelihood of the original ML tree was -5948732.6940. Using the HPC-PTHREADS-AVX architecture returned a tree with a log-likelihood of -5948749.6654, while IQ-TREE returned a tree with a log-likelihood of -5949077.382 (16.971 and 344.688 points worse than the RAxML v8.2.4 results, respectively).

Gene-wise log-likelihood comparative analysis across three candidate topologies

The results from the gene-wise comparisons of likelihood contributions showed that there were 155 orthologs in the data set that most strongly supported the MLT, 90 supported the RBT, and 142 supported the AT (Appendix A). The AT had a cumulative log-likelihood score that was >300 points worse, even though more individual genes support the AT over the RBT (Appendix A).

Edge-based comparative analyses across three candidate topologies

Of the three candidate topologies investigated by constraining key edges, Lecythidaceae sister to ESC received the best score, while Primulaceae sister to ESC received the worst (Appendix A). Similarly, the clade Lecythidaceae + Polemoniaceae + Fouquieriaceae + ESC received a better likelihood score than the clade Lecythidaceae + Primulaceae + ESC. Regarding the placement of Ebenaceae for this data set, Ebenaceae + Sapotaceae was supported by the highest number of orthologs whether or not two log-likelihood support difference was required (Appendix A). However, a number of orthologs support each of the investigated placements for Ebenaceae and less than half of genes supported any placement over the next best alternative by at least two log-likelihood points.

Gene duplication comparative analysis

Gene duplications mapped to each candidate backbone topology and the five additional hypothetical topologies revealed differing numbers of shared duplications that can be used as a metric of support among candidate topologies (Figure 2-2). Regarding which clade is better supported as sister to ESC, Lecythidaceae uniquely shared 433 duplications with ESC, more than twice as many as either alternative. Polemoniaceae + Fouquieriaceae shared 1300 unique duplications with Lecythidaceae + ESC, more than Primulaceae, which shared 626 unique duplications (Figure 2-2).

A corroborated topology for the 387 ortholog set

The above comparative analyses supported one topology among the three candidates identified for the 387 ortholog data set, namely the RBT (Figure 2-2; Appendix A). In this tree, all taxonomic families are recovered as monophyletic. Marcgraviaceae and Balsaminaceae (i.e., the balsaminoid Ericales) are sister, and form a clade that is sister to the rest of Ericales (i.e., the non-balsaminoid Ericales). Pentaphylacaceae is the earliest diverging clade within the Core Ericales. Ebenaceae and Sapotaceae form a clade that is sister to the Core. The monogeneric Fouquieriaceae is sister to Polemoniaceae. A clade containing Symplocaceae, Diapensiaceae, and Styracaceae is sister to Theaceae. Roridulaceae is sister to Actinidiaceae, and there was moderate support that this clade that is sister to Ericaceae. Sarraceniaceae, Roridulaceae, Actinidiaceae, and Ericaceae form a clade. A grade containing Primulaceae, Polemoniaceae + Fouquieriaceae, and Lecythidaceae leading to ESC is supported by the rapid bootstrapping, comparative duplication, and constraint-tree analyses.

Quartet sampling

We found varying levels of support for several key relationships in the RBT (Appendix **S9**). In our results and discussion we consider a QC score of (≥0.5) to be strong support because this signifies strong concordance among quartets (Pease et al., 2018). The monophyly for all families received strong support (QC ≥ 0.90). There was strong support (QC = 0.54) for the node placing Lecythidaceae sister to ESC, while equivocal support (QC = 0.035) for Polemoniaceae + Fouquieriaceae sister to Lecythidaceae and ESC. Within ESC, there was moderate support (QC = 0.28) for Ebenaceae sister to Sapotaceae, but poor (QC = –0.13) support for this clade as sister to the Core. There was strong support (QC = 0.85) for the clade including Symplocaceae, Diapensiaceae, and Styracaceae and moderate support (QC = 0.26) for this clade as sister to Theaceae. Roridulaceae was very strongly supported (QC = 0.99) as sister to Actinidiaceae but there was no support for Roridulaceae + Actinidiaceae + Ericaceae (QC = -

0.23). However, the monophyly of the clade that includes Roridulaceae, Actinidiaceae, Ericaceae, and Sarraceniaceae received very strong support (QC = 0.90).

The QD scores for several contentious relationships indicate that discordant quartets tended to be highly skewed towards one conflicting topology as indicated by scores below 0.3 (Pease et al., 2018). However, the QD score for the relationship placing Lecythidaceae sister to ESC in the RBT was 0.53, indicating relative equality in occurrence frequency of discordant topologies. Similarly, the relationship placing Polemoniaceae + Fouquieriaceae sister to Lecythidaceae + ESC received a QD score of 0.83, indicating that among alternative topologies (e.g., Primulaceae sister to the clade Lecythidaceae + ESC), there was no clear alternative to the RBT recovered through quartet sampling for this data set. The QI scores for all nodes defining interfamilial relationships were above 0.9, indicating that in the vast majority of sampled quartets there was a tree that was at least two log-likelihoods better than the alternatives. The QF scores for all but one taxon were above 0.70, and the majority were above 0.85, suggesting that rogue taxa were not a major issue (Pease et al., 2018).

Conflict analyses

Assessing ortholog tree concordance and conflict for the 387 ortholog set mapped to the RBT showed that backbone nodes were poorly supported with the majority of orthologs failing to achieve ≥70% bootstrap support (Appendix A). Among informative orthologs, the majority of trees conflict with any candidate topology at these nodes and there is no dominant alternative to the RBT. There were 77 ortholog trees with appropriate taxon sampling that placed Primulaceae sister to the rest of the non-balsaminoid Ericales, and 37 did so with at least 70% bootstrap support. The clade Lecythidaceae + Primulaceae + ESC was recovered in 47 ortholog trees, and in 17 with at least 70% bootstrap support. Of the 33 ortholog trees that contained the clade Primulaceae + Ebenaceae, nine did so with at least 70% bootstrap support.

Expanded ortholog sets

Seven combinations of relationships along the backbone were recovered in analyses of the expanded ortholog sets, which we term E-I though E-VII for reference (Figure 2-3). Among these, the ML tree estimated from the 2045-ortholog PRANK-aligned, partitioned supermatrix placed Lecythidaceae and Ebenaceae in a clade sister to Sapotaceae and the Core (E-II), with the other interfamilial relationships recapitulating those of the RBT. The topology recovered by ASTRAL for these orthologs (E-VII) placed Ebenaceae, Lecythidaceae, Polemoniaceae + Fouquieriaceae, and Primulaceae as successively sister to Sapotaceae and the Core. In regard to the placement of Ebenaceae for the 2045 PRANK-aligned set, the edge-based Phyckle analysis showed that 432 of orthologs in this data set with appropriate taxon sampling supported Ebenaceae sister to Primulaceae, while 202 did so by at least two log-likelihood over any alternative (Appendix A). However, 416 orthologs supported Ebenaceae sister to Sapotaceae (158 with ≥ 2LL), and 316 supported Ebenaceae sister to Lecythidaceae (140 with ≥ 2LL). When the 2045 ortholog set was aligned with MAFFT and concatenated into a supermatrix, the resulting ML topology (E-I) was such that the clade Primulaceae + Ebenaceae were sister to Sapotaceae, with that clade sister to the Core and Polemoniaceae + Fouquieriaceae sister to all of those. When ASTRAL was run with the 2045 MAFFT-aligned orthologs, the topology recovered (E-III) placed Polemoniaceae + Fouquieriaceae sister to the Core, with the clade Ebenaceae + Primulaceae and Lecythidaceae successively sister to those.

The backbone topology resulting from the 4682 ortholog PRANK-aligned, partitioned supermatrix was the same as that recovered with the 2045 ortholog and the same methods (E-II). The ASTRAL topology for the 4682 ortholog PRANK-aligned data set (E-VI) placed Lecythidaceae sister to Sapotaceae and the Core, with Primulaceae + Ebenaceae and Polemoniaceae + Fouquieriaceae successively sister to those. When each of the 449, 661, and 1899 ortholog sets were aligned with MAFFT to produce a supermatrix, the resulting ML backbone topology was the same as that of the MAFFT-aligned 2045 and 4682 ortholog sets (E-I), except when the 449 ortholog supermatrix was run without partitioning (E-II). When ortholog alignments were produced with PRANK, the ML backbone recovered from the 449 ortholog set was the same as that of the 2045 PRANK-aligned ML tree (E-II). The backbone of the ML trees produced from the 1899 and 661 PRANK-aligned ortholog sets were the same as that of the 2045 ortholog MAFFT-aligned ASTRAL tree (E-III). The 449, 661, and 1899 PRANK-aligned ortholog sets all produced the same backbone topology in ASTRAL (E-V). The backbone topologies recovered by ASTRAL for the 449, 661, and 1899 MAFFT-aligned ortholog sets also agree with one another (E-VI), but conflict with all other species trees recovered in this study.

Synthesis of uncertainty and determination of an overall consensus

In the species trees generated with the 387, 449, 661, 1899, 2045, and 4682 ortholog data sets, the relationships among several taxonomic families were in conflict with one another (Figure 2-3). Many of these relationships also conflict with the thoroughly investigated RBT, which was shown to be very well supported by the 387 ortholog set. Nine of ten ML trees generated with MAFFT-aligned supermatrices in the expanded data sets recovered Primulaceae and Ebenaceae sister to one another and forming a clade with Sapotaceae (E-I; Figure 2-3), however this pattern was not recovered under any other circumstances (Figure 1; Appendix A). In all, Primulaceae was recovered as sister to Ebenaceae in 17 of the 33 species trees generated in this study (51.5%), but these families were sister in only eight of the 24 trees (33.3%) where they did not form a clade with Sapotaceae. Sapotaceae was recovered as sister to the Core in 21 of the 33 species trees (63.6%). Thus, there is no majority consensus that reconciles these conflicting relationships; Primulaceae is only sister to Ebenaceae in a majority of trees if those that also contain the clade Sapotaceae + Ebenaceae + Primulaceae are considered and that topology is in direct conflict with Sapotaceae + Core, a relationship that is recovered in a majority of trees. In addition, edgewise support for a sister relationship between Primulaceae and Ebenaceae was data set dependent and showed equivocal gene-wise support. There was no majority for the phylogenetic placement of either Lecythidaceae or Polemoniaceae + Fouquieriaceae. Given this, the relationships among Primulaceae, Polemoniaceae and Fouquieriaceae, Lecythidaceae, Ebenaceae, and Sapotaceae are not resolved in the consensus topology. All other interfamilial relationships were unanimously supported by all species trees, except for the relationships between the clade Symplocaceae, Diapensiaceae, and Styracaceae sister to Theaceae, which was recovered in 29 of the 33 species trees (87.9%).

Gene and genome duplication inference

Gene duplications mapped to the consensus topology show that their largest number occur on the branch leading to the non-balsaminoid Ericales (Figure 2-4). Notable numbers of gene duplications also appear along branches leading to *Actinidia* (Actinidiaceae), *Camellia* (Theaceae), several members of Primulaceae, *Rhododendron* (Ericaceae), *Impatiens* (Balsaminaceae), and Polemoniaceae (i.e., *Phlox* and *Saltugilia* in this study). If gene duplications are mapped to the single most-commonly recovered species tree in this study (E-1), 3485 duplications occur along the branch leading to the non-balsaminoid Ericales (Appendix A). The K_s plots show peaks for several of the recent duplications, peaks between 0.1 and 0.5 (Appendix A). The K_s plots for most, but not all, taxa also contain a peak that appears between 0.8 and 1.5, as well as a peak between 2.0 and 2.5. The multispecies K_s plots for some ericalean taxa paired with a member of Cornales show an ortholog peak with a higher K_s value (i.e., farther to the right) than the paralog peaks near 1.0, suggesting two separate duplication events that each occurred after the divergence of the two orders. Some other combinations of ingroup and outgroup taxa resulted in an ortholog peak to the left of the paralogs peaks (Appendix A). When comparing the position of unambiguous ortholog peaks of all non-balsaminoid ericalean taxa to *Impatiens balsamifera*, the K_s value of the peak varied between values of 1.01–1.57 for *Schima superba* (Theaceae) and *Primula poissonii* (Primulaceae), respectively, indicating substantial rate heterogeneity in the accumulation of synonymous substitutions among ericalean taxa and implying that in the most extreme cases, some species have accumulated synonymous substitutions 55% faster than others (Appendix A).

Estimating substitutions supporting contentious clades

The estimated number of substitutions informing clades containing at least two families whose backbone placement is contentious tended to be much less than for relationships that garnered widespread support. These contentious clades had a median value of 8.65 estimated substitutions informing them, compared to 30.08, 75.09, and 144.00 informing the monophyly of Polemoniaceae and Fouquieriaceae, Lecythidaceae, and the non-balsaminoid Ericales, respectively, in the MAFFT-aligned ortholog trees (Figure 2-5). In trees for the same orthologs with alignments generated instead with PRANK, the corresponding median values were 8.96, 30.73, 74.16, and 145.98 respectively.

Discussion

Our focal data set, consisting of 387 orthologs, supports an evolutionary history of Ericales that is largely consistent with previous work on the clade for many, but not all, relationships (Appendix A). The balsaminoid clade, which includes the families Balsaminaceae, Tetrameristaceae (not sampled in this study), and Marcgraviaceae, was confidently recovered as sister to the rest of the order as has been shown previously (Geuten et al., 2004; Gitzendanner et al., 2018; Rose et al., 2018). Similarly, the monogeneric family Fouquieriaceae is sister to Polemoniaceae, the para-carnivorous Roridulaceae are sister to Actinidiaceae, and the circumscription of Primulaceae sensu lato is monophyletic (Rose et al., 2018). The majority of analyses for the 387 ortholog data set recovered a sister relationship between Sapotaceae and Ebenaceae, with that clade sister to the Core Ericales and Lecythidaceae sister to those. Notably, the topology supported suggests that Primulaceae diverged earlier than has been recovered in most previous phylogenetic studies and does not form a clade with Sapotaceae and Ebenaceae as has been suggested by others, including Rose et al. (2018). The topology recovered by Gitzendanner et al. (2018) using coding sequences from the chloroplast placed Primulaceae sister to Ebenaceae, with those as sister to the rest of the non-balsaminoid Ericales, though that study did not include sampling from Lecythidaceae. In addition to the traditional phylogenetic reconstruction methods, by applying the available data on gene duplications as a metric of support, and leveraging methods that make use of additional phylogenetic information present in the supermatrix, we were able to more holistically summarize the evidence present in a 387 ortholog data set in an effort to resolve the Ericales phylogeny (Figures 2-1 and 2-2; Appendix A).

It has been shown repeatedly that large phylogenetic data sets have a tendency to resolve relationships with strong support, even if the inferred topology is incorrect (Seo, 2008). However, some of our results suggest extreme sensitivity to tree-building methods. For example, the initial ML analysis resulted in an ML topology (MLT) with zero rapid bootstrap support for the placement of Lecythidaceae (Figure 2-1), while the rapid bootstrap consensus for this data set unanimously supported a conflicting relationship (i.e., Lecythidaceae sister to a clade including Ebenaceae, Sapotaceae, and the Core Ericales; the RBT). Gene-wise investigation of likelihood contribution confirmed that these two topologies had very similar likelihoods but did not identify outlier genes that seemed to have an outsized effect on ML calculation (Appendix A). Instead, the cumulative likelihood influence of the 387 genes in the supermatrix provides nearly equal support for the two topologies, while ASTRAL resulted in a third, and regular bootstrapping recovered an even more diverse set of topologies (Appendix A). These results suggest that there are several topologies with similar likelihood scores for this data set. Despite the fact that the additional comparative analyses applied to the 387 ortholog data set supported a single alternative among those investigated (i.e., the RBT), the recovery of multiple topologies by various tree-building and bootstrapping methods suggests that the criteria used to generate and filter orthologs could have marked potential to influence the outcome of our efforts to resolve relationships among the families of Ericales.

In addition to sensitivities associated with tree building methods, we investigated additional data sets constructed with a variety of methods and filtering parameters to shed further light on the nature of the problem of resolving the ericalean phylogeny. While it is clear from investigation of gene tree topologies for the 387 ortholog data set that phylogenetic conflict is the rule rather than the exception, the expanded data sets show that this is true for a variety of approaches to data set construction and not simply an artifact of one approach. While the monophyly of most major clades and the relationships discussed above were recovered across these data sets, we also demonstrate that many combinations of contentious backbone relationships can be recovered depending on the methods used in data set construction, alignment, and analysis (Figure 2-3).

An unresolved consensus topology

Based on the data available, we suggest that while the relationships recovered in the Core Ericales and within most families are robust across methodological alternatives, there is insufficient evidence to resolve several early-diverging relationships along the ericalean backbone. We therefore suggest that the appropriate representation, until further data collection efforts and analyses show otherwise, is as a polytomy (Figure 2-4). Whether this is biological or the result of data limitations remains to be determined. A biological polytomy (i.e., hard polytomy) can be the result of three or more lineages diverging rapidly without sufficient time for the accumulation of nucleotide substitutions or other genomic events to reconstruct the patterns of lineage divergence. Most of the inferred orthologs contain little information useful for inferring relationships along the backbone of the phylogeny; we investigated this explicitly by estimating the number of nucleotide substitutions that inform these backbone relationships and find that branches that would resolve the polytomy were based on a small fraction of the number of substitutions that informed better-supported clades (Figure 2-5). The phylogenetic signal presented in this study results in extensive gene tree conflict, albeit mostly with low support

(Appendix A). The major clades of Ericales may or may not have diverged simultaneously; however, if divergence occurred rapidly enough as to preclude the evolution of genomic synapomorphies, then a polytomy is a reasonable representation of such historical events rather than signifying a shortcoming in methodology or taxon sampling.

The high levels of gene tree conflict and lack of a clear consensus among data sets for a resolved topology is likely to have multiple causes. Among these is the fact that this series of divergence events seems to have happened relatively rapidly from about 110 to 100 million years ago (Rose et al., 2018). We also find evidence that a WGD is likely to have occurred before or during this ancient radiation (Figure 2-4; Appendix A); if this is the case, differential gene loss and retention during the process of diploidization is likely to complicate our ability to resolve the order of lineage divergences. In addition, we cannot exclude the possibility of ancient hybridization and introgression between these early lineages, because hybridization has been documented between plants that have been diverged for tens of millions of years (Arias et al., 2014; Rothfels et al., 2015). It is even possible that some of the lineages involved in such introgression have gone extinct in the intervening 100 million years, such that introgression from such now-extinct “ghost lineages” represent a insurmountable obstacle to fully understanding the events that lead to the diversity of forms we now find in Ericales. We chose not to test explicitly for evidence of hybridization here, because of the seeming equivocal phylogenetic signal present in most gene trees for these contentious relationships. We suggest that interpreting the generally weak signal present in most conflicting gene trees as anything other than a lack of reliable information, runs a high risk of overinterpreting these data because network analyses and tests for introgression generally treat gene tree topologies as fixed states known without error. However, future studies could potentially find such an approach to be appropriate for explaining the high levels of conflict among orthologs, but should carefully consider alternative explanations for gene tree discordance.

Gene and genome duplications in Ericales

Results from gene duplication analyses (Figure 2-6) showed evidence for several whole genome duplications in Ericales, including at least two—in *Camellia* and *Actinidia*—that have been verified using sequenced genomes (Huang et al., 2013; Shi et al., 2010; Xia et al., 2017). Our results strongly support the conclusion drawn by Wei et al. (2018) that the most recent WGD in *Camellia* is distinct from the *Ad*-α WGD that occurred in Actinidiaceae (Wei et al., 2018). We propose the name *Cm*-α for this WGD, which is shared by all *Camellia* in our study and may or may not also be shared by *Schima*, the only other genus in Theaceae that we sampled. Future studies with broader taxon sampling should be able determine whether the *Cm*-α WGD is shared by other genera in Theaceae or if it is exclusive to *Camellia*.

Our inferred genome duplications are concordant with several, but not all, of the conclusions drawn by Leebens-Mack et al. (2019), whose transcriptome assemblies comprised 24 of the 86 ingroup samples for this study. We find evidence for their ACCHα (i.e., *Ad*-α; Shi et al., 2010) and IMPAα WGDs in Actinidiaceae and Balsaminaceae, respectively, though our broader taxon sampling additionally reveals that both of these WGDs are shared by multiple species of their respective genera (Figure 2-4; Appendix A). Our results do not support their placement of ACCHβ, which would appear as a WGD shared exclusively by the Core Ericales in this study (Figure 2-4; Appendix A), nor do our results support the existence of DIOSα, which Leebens-Mack et al. (2019) infer to have occurred along a branch leading to a clade consisting of Polemoniaceae, Fouquieriaceae, Primulaceae, Sapotaceae, and Ebenaceae: a clade never recovered in this study, and recovered in only one of the three species tree methods employed by Leebens-Mack et al. (2019). We do not find evidence for MOUNα in Monotropoidiae and the in-paralog trimming procedure we employed precludes us from addressing the putative SOURα WGD because our sampling includes only one taxon from Marcgraviaceae (Leebens-Mack et al., 2019).

Our results suggest that a WGD occurred along the backbone of Ericales, either before or after the divergence of the balsaminoid clade, but after Ericales diverged with Cornales (Figure 2-4; Appendix A). Given the extent of the topological uncertainty recovered along the backbone of Ericales in this (Figure 2-3) and other studies (Gitzendanner et al., 2018; Leebens-Mack et al., 2019; Rose et al., 2018), and the fact that our single most-commonly recovered backbone (i.e., Topology E-I, Figure 3; Appendix A) would imply that most gene duplications occurred along the branch leading to the non-balsaminoid Ericales, we suggest that a single, shared WGD is the most justifiable explanation for the observed data, rather than a more complex series of nested WGD or near-simultaneous WGDs in sister lineages. We also infer notably high numbers of gene duplications along the branches within Ericaceae, Primulaceae, and Polemoniaceae, which suggests that these clades should be further investigated for evidence of novel, lineage-specific whole genome or other major chromosomal duplications (Figures 2-4 and 2-6).

The K_s plots for most of our ingroup species appear to share two peaks, in addition to peaks corresponding to several lineage-specific WGDs (Appendix A). One shared peak occurs between 2.0–2.5 in most taxa, which is often interpreted as corresponding to the genome duplication shared by all angiosperms (Jiao et al., 2012; Leebens-Mack et al., 2019). Our results show that this peak between 2.0 and 2.5 also includes to the "γ" palaeopolyploidization shared by the core Eudicots (Jiao et al., 2011). We are able to infer this by evaluating K_s plots for *Helianthus*, *Solanum*, and *Beta* (Appendix A). Because in-paralogs were trimmed in our homolog trees for all taxa, any WGDs in *Helianthus*, *Solanum*, or *Beta* not shared by another taxon in our study (i.e., Ericales and Cornales) will not appear in the K_s plot for that species. Therefore, the K_s plots for *Helianthus*, *Solanum*, and *Beta* will exclusively display evidence of polyploidization events that occurred before the MRCA of Asterales and Solanales in the cases of *Helianthus* and *Solanum*, or the MRCA of Asterales + Solanales and Caryophyllales in the case of *Beta* (Appendix A). Leebens-Mack et al. (2019) show that only polyploidizations at least as old as γ should be shared by these taxa and because none have a K_s peak with a value less than 2.0, that peak must include the γ event. Our characterization of the γ event is compatible with the conclusions of Qiao et al. (2019), who analyzed 141 sequenced genomes and found that the γ palaeopolyploidization corresponded to a $K_s$peak that ranged between 1.91 and 3.64 for 16 species that have not experienced a WGD since γ. Qiao et al. (2019) also fitted K_s distributions for their taxa with Gaussian mixture models; for *Actinidia chinensis*, fitted K_s peaks occurred at 0.317, 1.016, and 2.415, which correspond respectively to the first (*Ad*-α), second (*Ad*-β), and third (γ) most recent WGDs in *Actinidia* (Qiao et al., 2019).

Many of our ingroup species share a K_s peak occurring between 0.8 and 1.5 (Appendices A) that seems to correspond to a WGD shared by all non-balsaminoid Ericales (Figure 2-4; Appendix A). We suggest this is the *Ad*-β WGD characterized by Shi et al. (2010) and corroborated by Soza et al. (2019) and Qiao et al. (2019), the ACCHβ WGD recovered by Leebens-Mack et al. (2019), and the genome duplication Wei et al. (2018) concluded was shared between *Camellia* and *Actinidia* (Soza et al., 2019). Our study is the first with the necessary taxon sampling to show that the *Ad*-β WGD occurred in the ancestor of all or nearly all ericalean taxa and is likely shared by a clade that minimally includes Lecythidaceae, Polemoniaceae, Fouquieriaceae, Primulaceae, Ebenaceae, Sapotaceae, and the Core Ericales. The tree-based methods employed here precluded us from explicitly inferring the number of gene duplications that occurred directly before the divergence of the balsaminoid Ericales, because we employed a procedure that treated all non-ericalean taxa as outgroups for identifying duplicated ingroup clades in the homolog trees. Studies with broader taxonomic foci should investigate whether the balsaminoid clade share the *Ad*-β WGD with the rest of Ericales.

Our K_s plots are compatible with an uncharacterized WGD in *Rhododendron* (Ericaceae) as indicated by shared peaks near 0.5 in several taxa. Similarly, *Ardisia, Aegiceras*, and *Primula* (Primulaceae) also share a K_s peak near 0.5, compatible with a shared WGD in those taxa. Several taxa from *Phlox* and *Saltugilia* (Polemoniaceae) have a K_s peak near 0.12, though these peaks are relatively weak and do not provide strong support for a WGD. Future sampling of transcriptomes and genomes will likely lead to the discovery of additional, lineage-specific WGDs in Ericales and refine our understanding of which taxa share these and other duplication events (Yang et al., 2018).

Our results strongly suggest that uncertainty should be considered when inferring duplications with tree-based methods (Figure 2-2; Appendix A) because the species tree topology can determine where gene duplications appear to have occurred (Zwaenepoel and Van de Peer, 2019). The use of K_s plots as a second source of information may not completely ameliorate issues caused by topological uncertainty, because K_s plots are generally interpreted in the context of an accepted phylogeny (i.e., an error-free phylogeny where well-supported and contentious nodes are treated the same). Furthermore, K_s plots are affected by heterogeneity in evolutionary rate, with faster-evolving taxa accumulating synonymous substitutions at faster rates than more slowly evolving lineages (Appendix A), adding an additional complicating factor when comparing K_s peaks and synonymous ortholog divergence values across species (Qiao et al., 2019; Smith and Donoghue, 2008). Technical challenges such as missing data resulting from incomplete transcriptome sequencing, failure to assemble all paralogs in all gene families, biases in taxon sampling, as well as phylogenetic uncertainty in homolog trees, influences where many individual gene duplications appear in this and other studies of nonmodel organisms, and caution should be taken to avoid overinterpreting noisy signal as biological information.

Conclusions

The first transcriptomic data set broadly spanning Ericales and constructed from publicly available data resolves many of the relationships within the clade and supports several relationships that have been proposed previously. Our results confirm genome duplications in Actinidiaceae and Theaceae, and provide a more precise placement of a whole-genome duplication in an early ancestor of Ericales. We find evidence to suggest additional WGDs in Balsaminaceae, Ericaceae, Polemoniaceae, and Primulaceae. While our results were largely concordant within taxonomic families, the topological resolution of the deep divergences in Ericales is less decisive. We demonstrate that, with the available data, there is not enough information to strongly support any resolution, despite previous studies having considered these relationships resolved. Additional data will be needed to investigate the early divergences of the Ericales. Leveraging gene synteny and chromosome-level genome scaffolds could provide a promising direction for future attempts to resolve these relationships. Our analyses demonstrate that uncertainty needs to be thoroughly investigated in phylotranscriptomic data sets, because strong support can be given by different methods for conflicting topologies that can in turn affect the placement of WGDs on phylogenies. Even in a data set containing hundreds of genes and hundreds of thousands of characters, the criteria used in data set construction, as well as tree reconstruction methods and parameters, altered the inferred topology. Our results suggest that phylogenomic studies should employ a range of methodologies and support metrics so that topological uncertainty can be more fully explored and reported. The high prevalence of conflict among data sets and the lack of clear consensus in regards to the relationships among several major ericalean clades led us to conclude that a single, fully resolved tree is not supported by these transcriptome data, though we acknowledge that future improvements in sampling might justify their resolution.

Acknowledgements

The authors thank Christopher Dick, Greg Stull, Hannah Marx, Ning Wang, Caroline Edwards, members of the Smith lab, and two anonymous reviewers for their comments on an earlier draft of the manuscript. Support came from the National Science Foundation; D.A.L and J.F.W. were supported by NSF DEB 1207915, O.M.V. was supported by NSF FESD 1338694, and S.A.S. was supported by NSF DBI 1458466 and DEB 1917146.

Author Contributions

The study was conceived by D.A.L. and S.A.S. Data set construction was led by D.A.L. with contributions by O.M.V. and input from J.F.W. Analyses were led by D.A.L. and J.F.W with contributions and input by S.A.S. D.A.L. wrote the manuscript and produced the figures with contributions and input from O.M.V, J.F.W., and S.A.S. All authors read and approved the final draft of the manuscript.

Figures

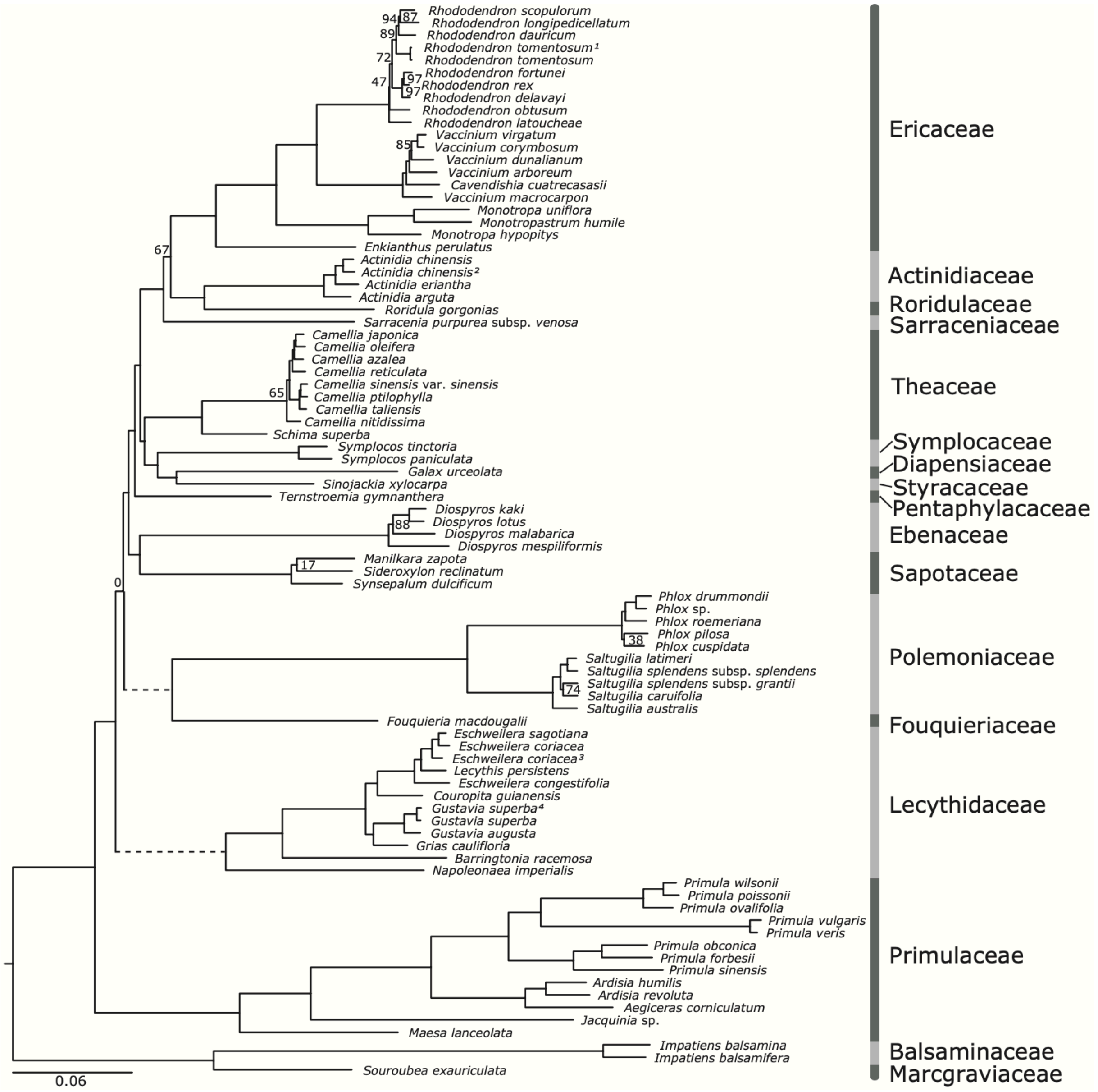


Figure 2-1. Maximum likelihood topology (MLT) recovered for a 387 ortholog dataset using RAxML. Nodes receiving less than 100% rapid BS support are labeled. Branch lengths are in substitutions per site. The node that determined the placement of Lecythidaceae received zero support (i.e. the MLT was never recovered by a rapid bootstrap replicate). Dashed lines indicate the two branches whose positions are transposed in the topology recovered by all rapid bootstrap replicates (i.e. the RBT).

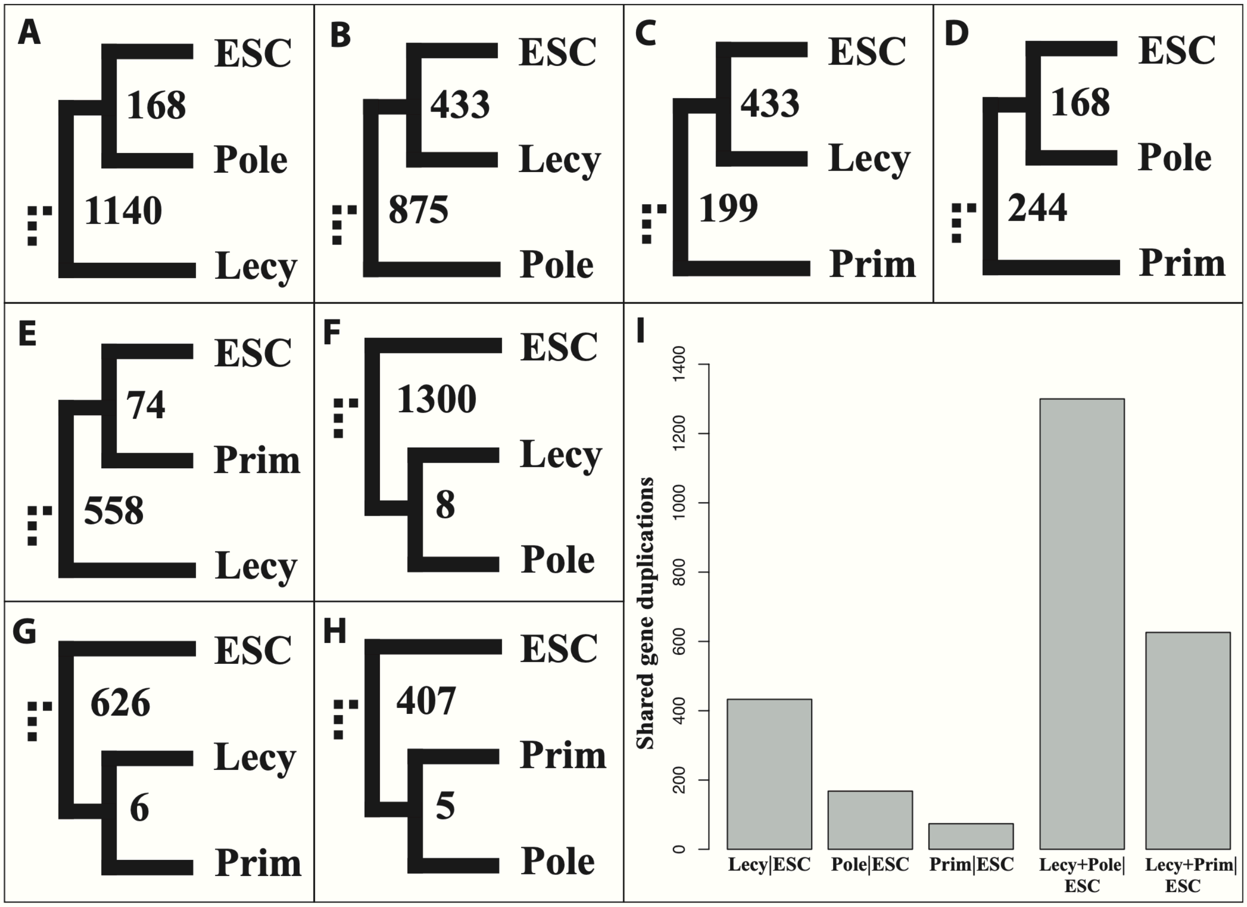


Figure 2-2. (A-H) Gene duplications with at least 50% SH-aRLT support in homolog trees mapped to several topologies recovered from the 387 ortholog dataset including the maximum likelihood topology (A), the rapid bootstrap topology (B), the ASTRAL topology (C), and several hypothetical topologies constructed to demonstrate evidence for shared duplications in clades not recovered with species tree methods (D-H). Names of clades are abbreviated to four letters, ESC represents the clade Ebenaceae + Sapotaceae + Core Ericales, and “Pole” represents the clade Polemoniaceae + Fouquieriaceae in all cases. I) Bar chart showing the number of uniquely shared gene duplications between clades that can be considered a metric of support for distinguishing among conflicting topological relationships.

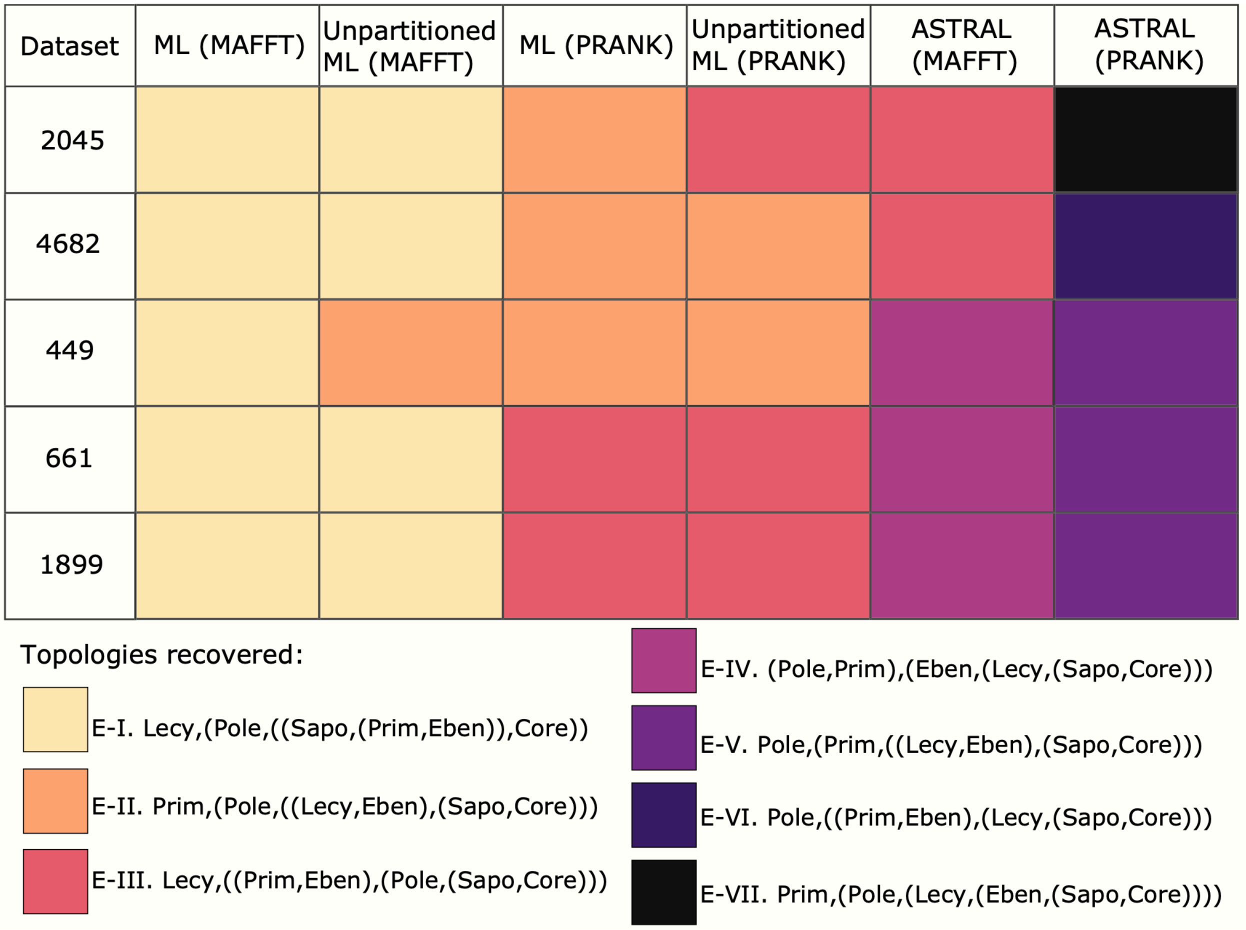


Figure 2-3. Topologies recovered from several combinations of ortholog datasets and species tree methods explored as alternatives to the focal 387 ortholog dataset. Each color corresponds to a unique backbone topology recovered in these analyses. Names of clades are abbreviated to four letters and "Pole" represents the clade Polemoniaceae + Fouquieriaceae in all cases.

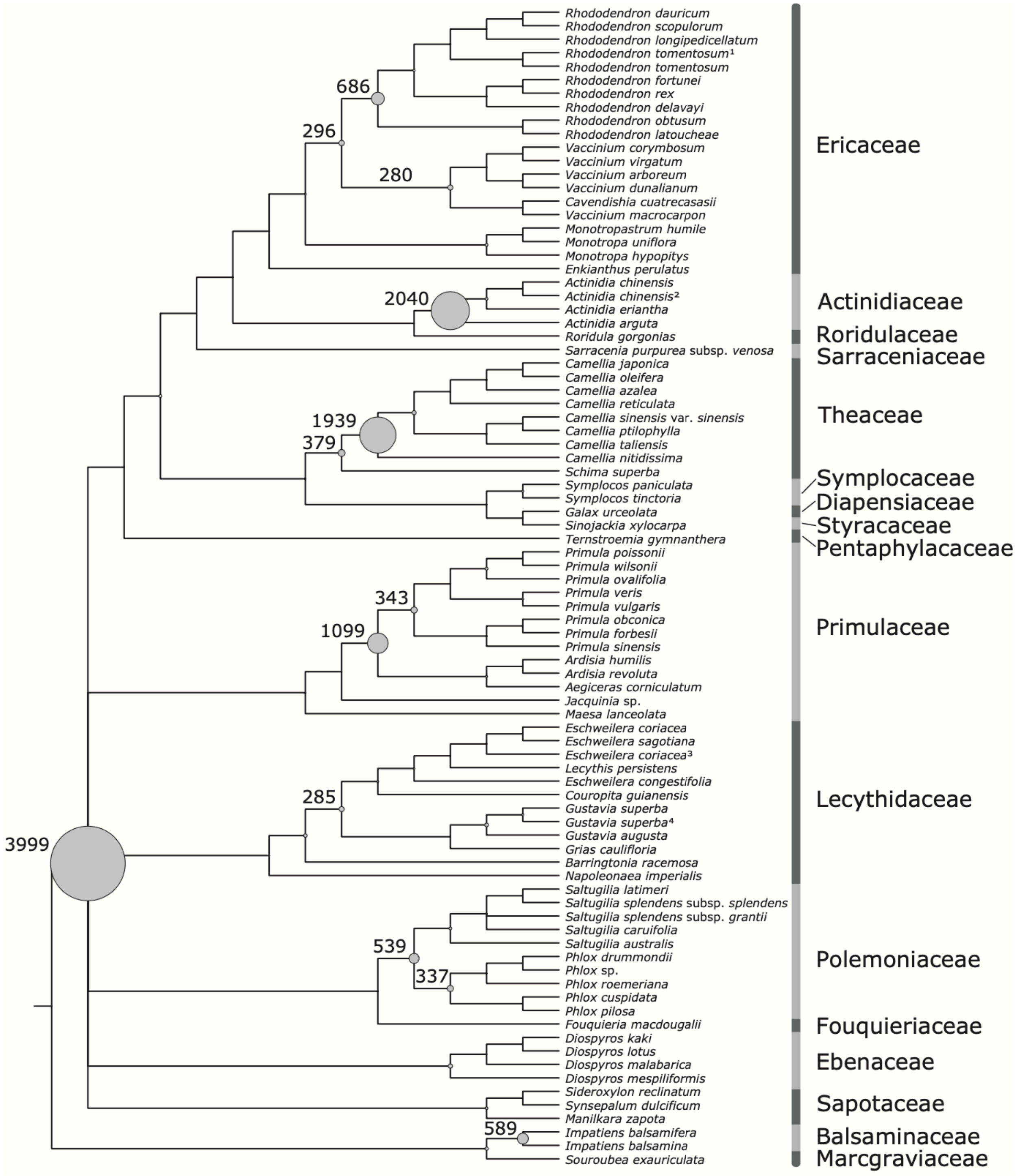


Figure 2-4. Gene duplications mapped to a cladogram of the consensus topology. Contentious relationships not supported by a plurality of methods were collapsed to a polytomy. The diameter of circles corresponds to the number of inferred duplications, and cases with at least 250 are labelled along branches. The single largest number of duplications occurs on the branch leading to the non-balsaminoid Ericales. Verified genome duplications in the Theaceae and the Actinidiaceae appear as the second and third largest numbers of inferred gene duplications respectively.

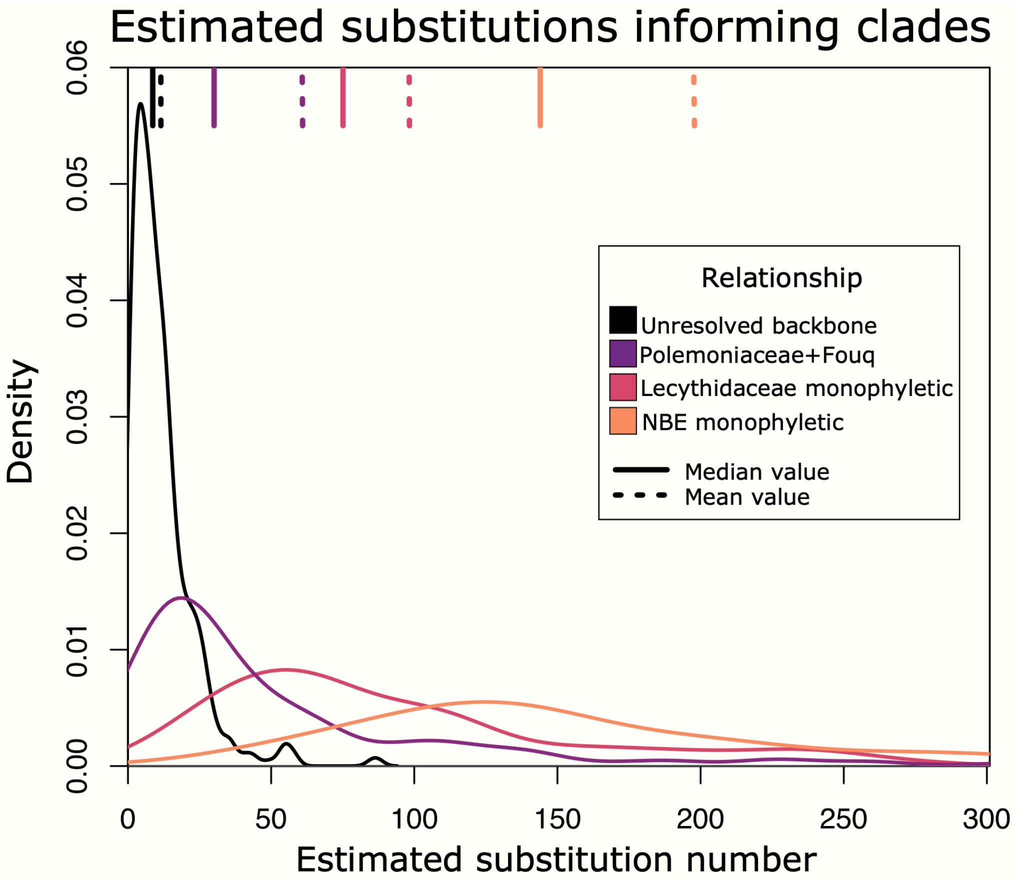


Figure 2-5. Estimated number of substitutions supporting clades using rooted orthologs from the 449 ortholog MAFFT-aligned dataset. The median number of estimated substitutions was 8.65 for clades that would resolve the polytomy in the consensus topology, compared to 30.08, 75.09, and 144.00 informing the monophyly of Polemoniaceae and Fouquieriaceae, Lecythidaceae, and the non-balsaminoid Ericales respectively.

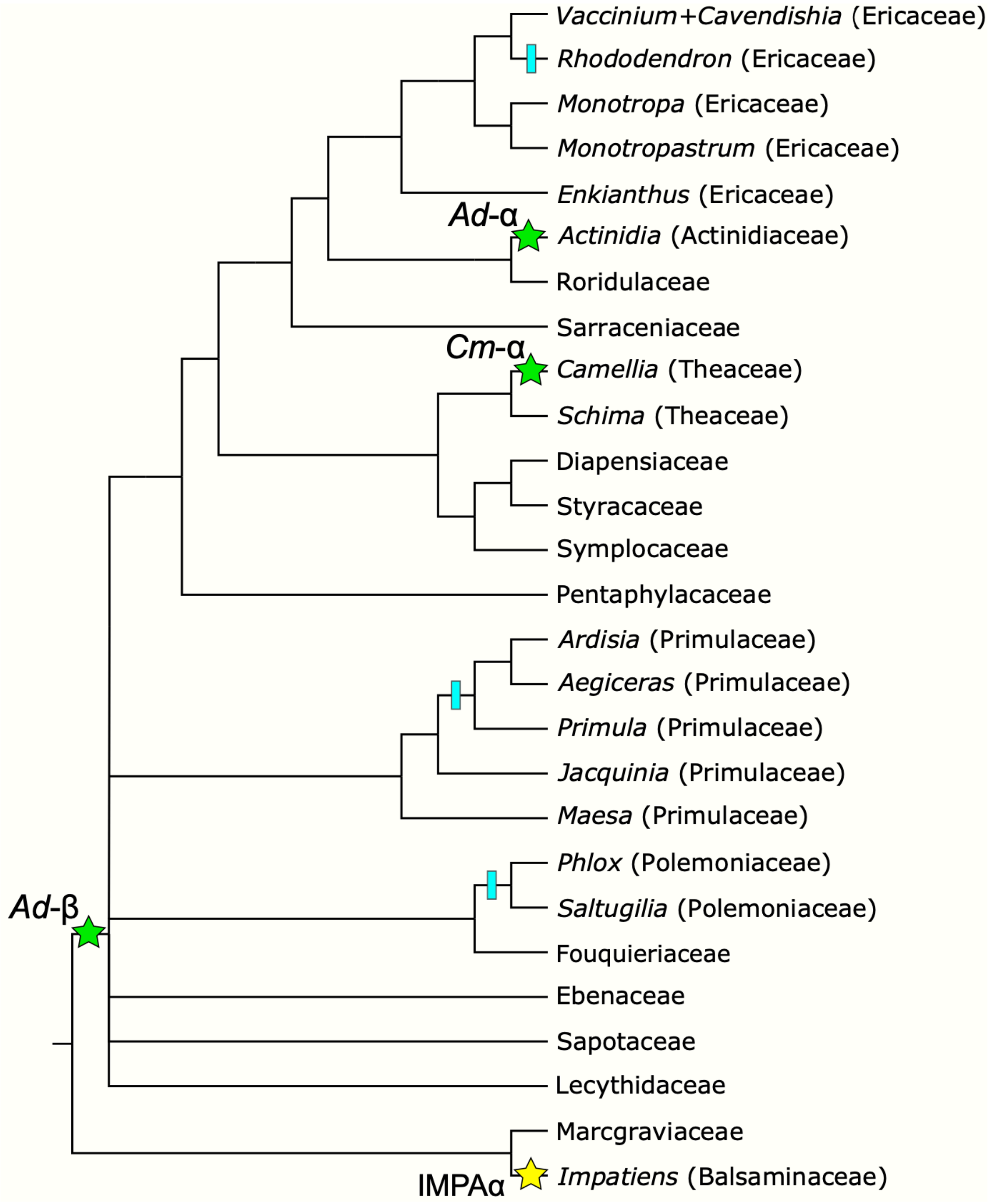


Figure 2-6. Putative whole genome duplications (WGDs) on a consensus phylogeny of Ericales. Placements are based on gene duplication analysis and K_s plots. Green stars represent WDGs that have been corroborated by sequenced genomes (Shi. et al., 2010; Xia et al., 2017; Soza et al., 2019). Yellow stars represent WGDs that have been proposed previously based on transcriptomes and are corroborated in this study (Leebens-Mack et al., 2019). Blue tick marks identify branches with evidence of previously uncharacterized WGDs that should be investigated in future studies. *Ad*-α and *Ad*-β are named after the WGD first detected in *Actinidia* (Shi et al. 2010). *Cm*-α is a name proposed in this study for a WGD unique to Theaceae that has been characterized previously and corroborated here (e.g. Wei et al., 2018). In cases where a possible or confirmed WGD was inferred along a branch leading to or within a botanical family, tips represent the genera sampled in this study. If no lineage-specific WGD was inferred for a family, the tip represents all taxa sampled for that family.

Chapter III

Admixture May Be Extensive Among Hyperdominant Amazon Rainforest Tree Species

Preamble: This chapter has been published in *New Phytologist*. The citation for this chapter is: Larson, D.A., Vargas, O.M., Vicentini, A., Dick, C.W., 2021. Admixture may be extensive among hyperdominant Amazon rainforest tree species. New Phytologist 232, 2520–2534.

Abstract

Admixture is a mechanism by which species of long-lived plants may acquire novel alleles. However, the potential role of admixture in the origin and maintenance of tropical plant diversity is unclear. We ask whether admixture occurs in an ecologically important clade of *Eschweilera* (Parvifolia clade, Lecythidaceae), which includes some of the most widespread and abundant tree species in Amazonian forests. Using target capture sequencing, we conducted a detailed phylogenomic investigation of 33 species in the Parvifolia clade and investigated specific hypotheses of admixture within a robust phylogenetic framework. We found strong evidence of admixture among three ecologically dominant species, *E. coriacea*, *E. wachenheimii*, and *E. parviflora*, but a lack of evidence for admixture among other lineages. Accepted species were largely distinguishable from one another, as was geographic structure within species. We show that hybridization may play a role in the evolution of the most widespread and ecologically variable Amazonian tree species. While admixture occurs among some species of *Eschweilera*, it

has not led to widespread erosion of most species' genetic or morphological identities. Therefore, current morphological based species circumscriptions appear to provide a useful characterization of the clade's lineage diversity.

Introduction

The extent to which hybridization and introgression (i.e. admixture) have affected the evolutionary history of tropical trees are only beginning to be understood. Admixture is expected to have various evolutionary consequences depending on the context of its occurrence, ranging from infrequent, localized production of hybrid offspring to the formation of new species (Rieseberg and Wendel, 1993). Adaptive introgression is a possible mechanism by which tropical tree populations may acquire favorable alleles, as has been demonstrated in various other plant clades (e.g., Leroy et al., 2020; Martin et al., 2006; Pease et al., 2016; Whitney et al., 2010), and may facilitate local adaptation beyond what might occur through selection acting on standing genetic variation and *de novo* mutations (Suarez-Gonzalez et al., 2018).

Hybridization among tropical trees has historically been considered a relatively rare phenomenon, primarily because of the dearth of morphological intermediates in herbarium specimens of tropical tree floras (Ashton, 1969; Parnell et al., 2013). However, recent work using next generation sequencing methods has demonstrated evidence of hybridization in tropical trees including in *Brownea* (Fabaceae; Schley et al., 2020), *Diosypyros* (Ebenaceae; Linan et al., 2020), *Melicope* (Rutaceae; Paetzold et al., 2019), and *Metrosideros* (Myrtaceae; Choi et al., 2020) among others. Caron et al. (2019) found that across tree taxa at a site in northern French Guiana, chloroplast haplotype diversity was more frequent in species with a local congener than those without, which they attribute to introgression. However, direct evidence of hybridization remains elusive for most clades of tropical trees. Because tests for admixture are inherently comparative, tests for admixture should ideally be nested within a robust and broadly inclusive phylogeny (Eaton et al., 2015). Such phylogenies are not yet available for many tropical clades, though phylogenomic datasets are becoming increasingly available (e.g., Christe et al., 2021; Couvreur et al., 2019; Linan et al., 2020; Loiseau et al., 2019; Prata et al., 2018). Investigations that characterize gene flow at well-studied forest plots may also enhance our understanding of the role of admixture in tropical forests, because 1) gene pools can be delimited without having to consider the confounding effects of geographic variation (Linan et al., 2020; Schley et al., 2020) and 2) permanently tagged trees provide a kind of "living herbarium" in which variation in field characters not evident in herbarium collections (e.g. branching architecture, microhabitat preferences, tree size) may be studied.

Target capture sequencing, also called target enrichment, is becoming increasingly popular for phylogenomic studies of non-model plants (Baker et al., 2021; Cronn et al., 2012) and often produces datasets with low missing data, even when the input DNA is partially degraded. The sizes of target loci vary, but generally range from hundreds to a few thousand base pairs (bp) in length. The number of targets also varies, but is frequently a few hundred loci, which is usually sufficient for phylogeny reconstruction but is far fewer than is typically used for inferring admixture, especially compared to methods such as RADseq, which can recover tens of thousands of RAD loci (Eaton et al., 2015; Eaton and Ree, 2013; Johnson et al., 2018; Vargas et al., 2020). Gene tree-based methods for inferring admixture using species networks can be used with several types of data, including target capture, though the resulting networks can include patterns of reticulate evolution that are sensitive to model parameters and gene tree quality (Morales-Briones et al., 2020). Given this and the increasing use of target capture for studies of plant evolution comes the need to explore additional methods capable of identifying evidence of admixture.

Our study taxa are tree species in the Brazil nut family, Lecythidaceae (Ericales). Lecythidaceae are ecologically important in many Neotropical forests and several species in the genus *Eschweilera* are among the most abundant trees across the Amazon basin (ter Steege et al., 2013). The Parvifolia clade of *Eschweilera* comprises 66 described species, characterized by morphological features including a distinctive double-coiled androecium (Figure 3-1D) and lateral arils on their seeds (Huang et al., 2015; Mori et al., 2010 onward). Several members of the Parvifolia clade have been described as hyperdominant (i.e. species with disproportionate abundance across a large area of the Amazon; ter Steege et al., 2013). The most abundant species of Lecythidaceae, *Eschweilera coriacea* (DC.) S.A.Mori, ranks third in abundance out of the more than 16,000 estimated Amazonian tree species. It is ecologically variable, thriving in floodplains as well as upland *terra firme* (Mori et al., 2010 onward), and is the only tree species that attains ecological dominance in all geographic subregions of the Amazon basin (ter Steege et al., 2020, 2013).

As is the case for many clades of tropical trees, species boundaries in Lecythidaceae are not precisely understood, though the taxonomy of the family is relatively well studied (e.g., Mori et al., 2010 onward; Mori and Prance, 1990; Prance and Mori, 1979). Previous studies have found discordance between morphology and plastid-based phylogenies, suggesting that chloroplast capture (i.e. the chloroplast of one species being introgressed into another) may be common in the group (Huang et al., 2015). However, hybridization followed by repeated directional backcrossing can result in chloroplast capture with little genetic or morphological evidence of nuclear admixture (Rieseberg and Soltis, 1991). A recent study using microsatellite

DNA markers suggested that the nuclear genomes of *Eschweilera* may also conflict with morphological based species circumscriptions (Heuertz et al., 2020), though we are not aware of any previous studies that have shown explicit evidence of nuclear admixture in Lecythidaceae.

We addressed the following questions: 1) is there evidence of nuclear admixture among species of the Parvifolia clade of *Eschweilera*, including species that are among the most abundant and ecologically variable trees in the Neotropics? and 2) to what extent do accepted species of *Eschweilera* represent monophyletic lineages that are distinguishable from one another using nuclear genomic data? The answers to these questions may shed light on whether the hyperabundance of widespread species like *E. coriacea* could be partly explained by a history of genetic introgression. We employed a multi-faceted sampling strategy and used target capture sequencing to generate the largest phylogenomic dataset for the family to date. Our methods included the implementation of an explicit test for admixture suitable for target capture data, which may prove useful for other phylogenomic datasets.

Materials and Methods

Focal study site and sampling strategy

We conducted sampling using two approaches. First, we sampled 12 focal species of the Parvifolia clade (Table 3-1; Appendix B) that co-occur at a single 100-ha forest plot in which all individuals of Lecythidaceae ≥ 10 cm diameter at 1.3m height have been tagged and identified by specialists beginning in the late 1980s (Mori et al., 2001; Mori and Lepsch-Cunha, 1995). The "Lecythidaceae plot" lies within Reserve 1501, also known as Km 41, of the Biological Dynamics of Forest Fragments Project (BDFFP) located approximately 70 km north of Manaus, Brazil (2° 24' 54" S, 59° 50' 39" W). The plot was established to study the Lecythidaceae of the central Amazon, a geographic center of diversity for the clade, but an area in which its taxonomy was poorly characterized (Mori and Lepsch-Cunha, 1995). By pairing ecological studies with alpha taxonomy, the investigators sought to characterize nuanced morphological differences among species across population samples and, in doing so, identify new species and their ecological differences (Mori et al., 2001; Mori and Lepsch-Cunha, 1995). Flowers and fruits are produced only sporadically in many species of Lecythidaceae, but species determinations for each tree in the plot were made using fertile material whenever possible (Mori et al., 2001). The site was re-censused in 2019, which showed there to be 6741 trees from 36 described species of Lecythidaceae (Milton et al., 2022). Herein, we refer to this 100-ha Lecythidaceae plot as Reserve 1501.

We chose focal species that were among the most abundant and most closely related species of Lecythidaceae at Reserve 1501 (Huang et al., 2015; Milton et al., 2022). Whenever possible, we sampled four to six tagged trees of each focal species and observed a minimum of at least 100 meters between conspecifics to reduce the chances of sampling immediate relatives. Our field collections relied on prior tree identifications of S. Mori and coworkers and we prioritized collection of three individual trees that seemed to have intermediate morphology, including in branching architecture (Appendix B). For each field-collected sample, leaf tissue was desiccated in silica gel and a voucher was deposited at the BDFFP collection at the National Institute of Amazonian Research (INPA), in Manaus, Brazil. In total, our sampling included 60 individuals collected at Reserve 1501 that were identified as a focal species or suspected hybrid based on morphology (Table 3-1).

Our second sampling approach aimed for wider phylogenetic and geographic breadth and used herbarium material and existing forest inventory vouchers. For this broader sampling, the

New York Botanical Garden Herbarium (NY) provided about half of our samples, which also included several non-focal species collected at Reserve 1501 and the surrounding area. Our overall sampling included 240 individuals from 127 of the 230 described species of Neotropical Lecythidaceae and seven outgroup species from Paleotropical genera. This included 109 individuals of the Parvifolia clade from 33 described species as well as several species that have not yet been formally described (Appendix B). A full analysis of the relationships among all major clades, as well as a revised taxonomy of Lecythidaceae utilizing this sampling is forthcoming (O. Vargas et al., *in prep*).

Sequencing and assembly

We performed DNA extractions using the NucleoSpin Plant Mini Kit II (Macherey-Nagel, Düren, Germany) following the manufacturer's protocol, but we extended the digestion step to one hour and added 5 uL of proteinase K (20 mg/mL; Qiagen, Hilden, Germany). Preparation of unenriched libraries for genome skimming and target-enriched libraries followed by 150 bp paired-end sequencing on an Illumina HiSeq4000 machine (Illumina Inc., San Diego California, USA) was performed by Rapid Genomics (Gainesville, Florida, USA). The probes used to enrich libraries were designed to capture 344 nuclear genes previously inferred to be low or single copy and genetically variable in Lecythidaceae (Vargas et al., 2019). Raw reads were processed with SeqyClean (Zhbannikov et al., 2017) to trim sequencing adapters, filter out low-quality reads, and trim read sections with a Phred score < 20 using a window of 10 bp. Trimmed reads were checked with FASTQC v0.11.3 (Andrews, 2010). Target loci were assembled using HybPiper v1.3.1 (Johnson et al., 2016) with default settings and a target file that included DNA sequences based on complete cDNA targets (Vargas et al., 2019). The Hybpiper pipeline uses

Exonerate (Slater and Birney, 2005), BLAST+ (Camacho et al., 2009), Biopython (Cock et al., 2009), BWA (Li and Durbin, 2009), SAMtools (Li et al., 2009), GNU Parallel (Tange, 2011), and SPAdes (Bankevich et al., 2012).

Paralog filtering and alignment

When employing target capture, paralogs can be enriched during library preparation and recovered in locus assemblies. While evidence suggests all or most Lecythidaceae are diploid (Heuertz et al., 2020), the lineage is thought to have experienced a whole genome duplication that occurred near the time of the most recent common ancestor of Ericales (Larson et al., 2020 [Chapter II of this dissertation]). Given that paralogs from gene duplications can confound many phylogenetic analyses, we employed a tree-based pruning approach meant to reduce misidentified orthologs and assembly errors (Yang and Smith, 2014). The parameters used in this trimming procedure were derived based *a priori* knowledge of the Lecythidaceae phylogeny and inspection of hundreds of amino acid phylogenies (Appendix B; Mori et al., 2015; Rose et al., 2018; Larson et al., 2020). The procedure included multiple sequence alignment with MAFFT v7.271 (Katoh et al., 2002; Katoh and Standley, 2013) followed by amino acid tree estimation with RAxML v8.2.11 (Stamatakis, 2014) and was meant to reduce non-orthologous sequences in the orthogroup alignments, while minimizing loss of phylogenetic information for taxa in the Parvifolia clade (Appendix B). We use the term orthogroup to denote groups of sequences that appear to be reciprocally orthologous based on sequence similarity, regardless of their present function in individual species.

Preliminary phylogenetic investigation

In order to identify clades of closely related individuals, check determinations for specimens, and to verify which individuals were nested within the Parvifolia clade, a phylogenetic tree (herein referred to as the preliminary phylogeny) was estimated with the assembled sequences from all 240 samples after the paralog filtering procedure described above. The preliminary phylogeny was estimated using RAxML v8.2.11 and a separate GTRCAT model partition for each of the exon and intron alignments of each orthogroup (Stamatakis, 2014). To assess support for clades in the preliminary phylogeny, rapid bootstrapping with 200 replicates was conducted. The results were visualized with Figtree (https://github.com/rambaut/figtree/).

Genotyping and SNP analysis

In order to investigate the genetic structure of Parvifolia species and identify potentially admixed individuals, we called SNPs for each individual using GATK v.4.1.0.0 (McKenna et al., 2010). The exon sequences for one individual for which we recovered 343 target loci with a combined length of 836,403 bp were used as a reference assembly (Appendix B). Genomic variants were called for each individual following GATK best practices, with modifications where necessary (Appendix B) to accommodate the available genomic resources for these non-model species (DePristo et al., 2011; Hanlon et al., 2019; Li, 2013; Li et al., 2009; Poplin et al., 2017; Van der Auwera et al., 2013). Several clades were identified based on the preliminary phylogeny and a clade-specific SNP dataset in approximate linkage equilibrium (Appendix B) was generated for each (Purcell et al., 2007). We used *Structure* v2.3.4 (Pritchard et al., 2000) to investigate genetic clustering of individuals within each clade and determined the most appropriate number of populations (K) for each subset of taxa by comparing the estimated posterior probability of the data for multiple values of K in conjunction with *a priori* taxonomic information (Appendix B). In cases where an individual showed strong evidence of clustering with a species other than that to which it was identified based on morphology, the identity of the individual was further investigated, and its determination was updated to reflect taxonomic uncertainty and all available evidence (Appendix B). Special consideration was given to *E. roseocalyx* (Batista et al., 2017), which appeared to be nested within the broadly distributed species *E. coriacea* based on preliminary results (Appendix B). To further explore patterns of genetic variation within *E. coriacea*, we used the gdsfmt and SNPRelate packages (Zheng et al., 2012) in R v3.6.0 (R Core Team, 2019) to produce an additional SNP dataset and conducted a genetic principal component analysis (PCA), which was visualized with a custom R script that utilized the *plotly.js* library (Sievert, 2020).

Verifying admixture with rooted triple tests

To corroborate the admixed ancestry of individuals identified using *Structure* and test for evidence of ancestral introgression among closely related species, we implemented a test capable of inferring admixture from a set of gene trees using rooted triplets (RT; i.e. gene trees consisting of three ingroup individuals and an outgroup; Figure 3-2), which we conducted using the novel script *Run_RT_tests.py* (see Data Availability Statement). A version of this test has been proposed previously (Huson et al., 2005), but we are not aware of any previous studies that have used it to investigate admixture in target capture datasets. The RT tests were conducted by subsetting each orthogroup alignment to include the four individuals of interest and estimating a gene tree with branch lengths for each sub-alignment using IQ-TREE v1.6.3 (Chernomor et al., 2016; Nguyen et al., 2015). This obviated the need to re-align sequences for each test and allowed the sequence data from all 240 samples to inform the sub-alignment, which may have helped alleviate alignment issues due to missing data. Then, the topologies of the resulting trees were summarized to assess whether the data were compatible with a scenario of no admixture, using the same theoretical framework as the *D*-statistic (Green et al., 2010). However, unlike most implementations of the *D*-statistic that count patterns in multiple sequence alignments or SNP datasets, our test is based on gene trees and can therefore readily be used with phylogenomic datasets consisting of relatively large gene regions in which all sites within a region are assumed to share the same phylogenetic history.

When all four-taxon gene trees are rooted on a known outgroup, the result is a set of rooted triplets, each of which contains exactly one ingroup relationship. There is a single tree bipartition that contains topological information for the ingroup, since two individuals will be sister to the exclusion of the third. For a rooted triplet consisting of ingroup taxa A, B, and C, the three possible ingroup bipartitions are (AB|C), (AC|B), and (BC|A). We define the most frequent of the three bipartitions as the "major relationship" and the other two possibilities as "conflicting relationships". The two individuals that form the major relationship are inferred to be the two that are most closely related and are herein referred to as T_1 and T_2 (Figure 3-2). T_1 and T_2 are assumed to share a most recent common ancestor (MRCA) that occurred more recently than the MRCA of all three ingroup individuals, whether or not there is ongoing gene flow between/within the population(s) to which T_1 and T_2 belong (i.e. they can be the same or different species). As long as there is a null expectation of no gene flow with the populations to which the third ingroup (T_3) or the outgroup (O) individuals belong (i.e. T_3 and O are different species from one another as well as from T_1 and T_2) and it can be assumed that for each gene tree, O has the earliest diverging sequence, then in the absence of gene flow between the lineages represented by T_3 and T_1 and/or T_2, the number of gene tree with each of the possible two conflicting relationships should be statistically equal (Bryant and Hahn, 2020), because each is equally likely to occur due to incomplete lineage sorting (ILS).

Any statistically significant deviation from equality can be considered evidence that the assumptions of the multispecies coalescent model have been violated by gene flow between the lineages to which T_3 and T_1 and/or T_2 belong. We calculate *P* as the probability of a result at least as unequal as the observed frequencies using a binomial test where each gene tree that conflicts with the major relationship represents a trial and the probability of either conflicting relationship is equal to 0.5.

To correct for multiple comparisons, we used the Holm-Bonferroni method with α=0.01 to adjust our critical value for rejecting the null hypothesis (Eaton et al., 2015; Holm, 1979). The statistical power of each RT test is affected by the number of gene trees that conflict with the major relationship, which is expected to vary based on the time since the MCRA of the relevant individuals. The type II error rate (i.e. failing to reject the null hypothesis of no admixture when in fact there has been admixture) of this type of RT test may be relatively high for many target capture datasets, due to the relatively low number of independent trials available compared to some other tests for admixture using RADseq or whole genome assemblies. Because of this, our results may represent a conservative estimate of admixture among our sampled species, especially for cases of historical introgression involving small proportions of the genome. However, our statically significant results provide strong evidence of admixture.

It should be noted that because we utilized coding sequences and the introns adjacent to them, each locus is subject to natural selection. However, it is unlikely that selection would generally lead to a systematic bias for one conflicting gene tree topology over the other for a large enough number of independent loci to significantly increase the type I error rate (i.e. rejecting the null hypothesis of no admixture, when in fact no admixture has occurred). It is also important to note that the test as implemented does not explicitly account for heterozygosity, since each locus is represented by a single consensus sequence per sample, as is typical in most phylogenomic datasets. The effect that differing consensus-calling approaches during sequence assembly might have on phylogeny-based inferences of admixture warrants future study.

Parvifolia clade phylogeny

To build a robust phylogenetic hypothesis for the Parvifolia clade, we conducted additional analyses without individuals with evidence of recent admixture. We used additional tree-based paralog pruning and generated two supermatrices, one that included data from introns and another that did not (Appendix B). For clarity, we refer to the best-scoring tree for the dataset that included both intron and exon sequences as the "Parvifolia phylogeny" and the best-scoring tree for the other supermatrix as the "exon-only Parvifolia phylogeny". For visualization, a version of each phylogeny was produced by trimming tips to include a single representative of each accepted species (Appendix B) using the pxrmt function in phyx (Brown et al., 2017a). Conflict between the reduced-representation phylogenies was visualized using the phytools package in R (Revell, 2012). A version of the Parvifolia phylogeny with all tips, as well as an analysis of topological conflict with the untrimmed exon-only Parvifolia phylogeny, generated using the pxbp function in phyx, is also reported.

Summaries of collection records and phenology for selected species

In order to visualize the extent of known range overlap among hyperdominant species *E. coriacea*, *E. parviflora*, *E. truncata*, and *E. wachenheimii*, we used a dataset curated by Mori et al. (2017) comprising available species occurrence records for these taxa (Vargas and Dick, 2020). All records for each species were plotted with QGIS v3.16.3 (https://github.com/qgis). We used a river shapefile available from the World Bank (https://datacatalog.worldbank.org/dataset/major-rivers-world, CC-BY 4.0 license), the World Borders Dataset (http://thematicmapping.org, CC BY-SA 3.0 license), and an digital elevation model (Lehner and Grill, 2013). We plotted individual occurrences, rather than range summaries, to more clearly show the available data and corresponding gaps in existing collection records. To investigate flowering times of *E. coriacea*, *E. parviflora*, and *E. wachenheimii*, we used the C.V. Starr Virtual Herbarium (http://sweetgum.nybg.org/science/vh/) to examine all collections from Amazonas, Brazil housed at NY. We identified specimens with flowers or flower buds at time of collection and verified the collection date and determination for each based on the specimen label. The results were plotted as box plots and dot plots for each species in R using ggplot2 (Wickham, 2016) after removing duplicate collections made from the same tree on the same day.

Results

Admixture among species of the Parvifolia clade

Our SNP-calling approach identified 148,310 polymorphic sites among 109 individuals in the Parvifolia clade. Both *Structure* analyses and RT tests support evidence of admixture among two species pairs in our sampling. Two individuals collected at Reserve 1501 were supported as having near equal ancestry of *E. coriacea* and *E. wachenheimii* (Figure 3-3). These individuals were not recovered as sister to one another in the preliminary phylogeny (Appendix

B) and RT tests showed significant evidence of admixture for separate tests that included these individuals (Table 3-2; Appendix B). Two additional individuals were supported as genetic intermediates between *E. wachenheimii* (ca. 70-75% ancestry) and *E. parviflora* (ca. 25-30% ancestry) in *Structure*, with RT tests also supporting evidence of admixture (Figure 3-3; Table 3-2; Appendix B). This second pair of individuals were recovered as sister to one another in the preliminary phylogeny (Appendix B).

We also tested for evidence of more ancient introgression among lineages using RT tests with three ingroup individuals from three different species determinations or *Structure* clusters (in cases where the individual's identity was unclear). Individuals whose determination contained an *affinis* modifier were considered to be their own lineage for this purpose. We conducted 25 such tests, selecting one individual per lineage and excluding individuals with evidence of recent admixture in *Structure* analyses. We did not find significant evidence of admixture in any of these tests (Figure 3-4; Table 3-2; Appendix B), though three resulted in an uncorrected $P<0.05$ but that was not significant at the level of $\alpha=0.01$ after correcting for multiple tests with the Holm-Bonferroni method (Table 3-2; Appendix B). One such test included individuals determined as *E. parviflora*, *E.* aff. *parviflora*, and *E. wachenheimii* in which 63.3% of conflicting gene trees supported one alternative ($P = 3.52\times10^{-3}$). Another test included individuals of *E. laevicarpa*, *E. bracteosa* and an individual determined as *E.* aff. *laevicarpa*: for this test 59.4% of conflicting gene trees supporting one alternative ($P = 8.58\times10^{-3}$). The third test, which included individuals of *E. truncata*, *E. coriacea*, and *E. sagotiana*, resulted in 58.8% of conflicting gene trees supporting one alternative ($P = 1.19\times10^{-2}$; Table 3-2; Appendix B).

Monophyly of described species in the Parvifolia clade

In *Structure* analyses, individuals collected at Reserve 1501 were consistently assigned ancestry corresponding almost exclusively (i.e. greater than 95%) to a single cluster, with notable expectations for two individuals with evidence of admixture (Figures 3-3 and 3-4; Appendix B). There did not appear to be admixture within several clades based on samples collected at Reserve 1501 including 1) *E. collina*, *E. bracteosa*, and *E. laevicarpa*, 2) *E. atropetiolata* and *E. cyathiformis*, or 3) *E. micrantha* and *E. rankiniae* (Figure 3-4). When considering individuals from these species collected outside our focal plot, some were inferred to have ancestry corresponding to multiple species. However, this appeared to be the result of intraspecific variation due to geographic structure, as there was no evidence of admixture in relevant RT tests (Appendix B). Intraspecific variation could have caused ancestry to be assigned to a second cluster due to the parameterization of the analysis or uneven sampling across subpopulations (e.g., several individuals sampled from Reserve 1501, one individual from another locality). Indeed, the tendency for *Structure* to assign mixed ancestry in the presence of isolation by distance (Pritchard et al., 2010) or when sampling is uneven across hierarchical levels of population structure (Puechmaille, 2016) has been well-documented. Alternatively, this signal could represent admixture that RT tests failed to detect.

Overall, most individuals had morphological determinations that agreed with genetic evidence. There were 60 individuals collected at Reserve 1501 with morphological determinations as one of our focal species or suspected hybrids. Of these, seven (11.7%) were shown to require redeterminations based on genetic evidence and two were shown to be admixed (Appendix B). There were 51 individuals in our broader sampling of the Parvifolia clade that did not meet both of the following criteria: 1) determined to be a focal species based on morphology;

and 2) collected at Reserve 1501. Of these 51, there were 11 (21.6%) that required redeterminations, and two that showed evidence of admixture. Seven could be redetermined to species and four were assigned a putative species determination with an *affinis* modifier to reflect uncertainty (Appendix B).

Geographic structure in E. coriacea

There was strong evidence of geographic structure among 12 samples of *E. coriacea* with no evidence of recent admixture. In a PCA of SNP data, the first, second, and third principal components explained 15.14%, 11.24%, and 10.82% of the total variance respectively and individuals with the same country of origin clustered together (Appendix B). In phylogenetic analyses, collections from Brazil formed a clade which was strongly supported as sister to collections from French Guiana (Appendix B). The single individual collected in Panama was sister to an individual collected at Los Amigos field station at Madre de Dios, Peru, with those sister to a clade of two individuals collected at Yasuní National Park in Ecuador; those four individuals were also inferred to have varying amounts of ancestry corresponding to a second cluster in *Structure* analyses, while individuals from Brazil and French Guiana had inferred ancestry almost exclusively corresponding to a single cluster (Figure 3; Appendix B).

Phylogenetic relationships in the Parvifolia clade

Our target capture approach resolved most of the phylogenetic relationships among sampled species of the Parvifolia clade, though for some, support was dataset-dependent (Figure 3-4; Appendix B). Seven relationships among accepted species differed between the Parvifolia phylogeny (i.e. intron and exon data) and the exon-only Parvifolia phylogeny (Appendix B).

Inferred relationships among individuals within a species tended to vary more than relationships among species across datasets (Appendix B). Regardless of whether intron data was included, *E. truncata* and *E. wachenheimii* were inferred to be sister taxa, as were *E. coriacea* and *E. parviflora*. Those four species formed a clade with *E. sagotiana*, with that clade of five species sister to a clade consisting of *E. pedicellata*, *E. ovata*, and *E. albiflora* (Figure 3-4).

Summary of collection records and phenology of selected species

Existing collection records showed broad overlap in the geographic ranges of the four species we investigated (Figure 3-5). Our survey of phenology yielded 63 unique collections in flower from Amazonas, Brazil (Appendix B). Collection date ranges and interquartile ranges for each of the three species overlapped, with the medians for each falling within three weeks of one another during the dry season (Appendix B).

Discussion

Admixture in the Parvifolia clade

Our results add to the small but growing body of evidence regarding admixture among tropical trees and are, to our knowledge, the first examples of nuclear admixture among hyperdominant Amazonian species. Our sampling included all accepted species of the Parvifolia clade known to occur in the intensively studied Reserve 1501 plot (Mori and Lepsch-Cunha, 1995). All individuals of our 12 focal species collected at Reserve 1501 could be assigned robust species determinations based on *Structure* analyses and tree-based phylogenomic inference (Figures 3-3 and 3-4; Appendix B). Our results provide robust evidence of admixture between two of our focal species, *E. coriacea* and *E. wachenheimii*. The two *E. coriacea* × *wachenheimii* individuals were recovered as successively sister to all *E. wachenheimii* individuals in our preliminary phylogeny, consistent with each sharing a high degree of genetic similarity with *E. wachenheimii* while also harboring genetic dissimilarities with *E. wachenheimii* and with one another (Appendix B). In addition, there was significant evidence for rejecting the null hypothesis of no admixture for RT tests that included one *E. wachenheimii*, one *E. coriacea*, and either putative *E. coriacea* × *wachenheimii* individual (Table 3-2; Appendix B).

Our results also strongly support hypotheses of admixture between *E. wachenheimii* and *E. parviflora* (Figure 3-3; Table 3-2). The *E. parviflora* × *wachenheimii* individuals were inferred to have unequal ancestry from the two parent species, suggesting that hybridization followed by backcrossing may have occurred (Figure 3-3; Appendix B). We note that only a single individual of *E. parviflora* has ever been recorded at Reserve 1501 and therefore was not among our focal species; the collections of these admixed individuals were made within the BR-319 plot network, south of Reserve 1501 (https://ppbio.inpa.gov.br/sitios/br319; Appendix B). Both sampled individuals with >95% ancestry corresponding to *E. parviflora* in *Structure* analyses were collected in French Guiana. However, our results are not consistent with geographic structure: the relevant RT tests rejected the null hypothesis of no admixture for triplets consisting of one *E. wachenheimii*, one *E. parviflora*, and either putative *E. parviflora* × *wachenheimii* intermediate (Table 3-2; Appendix B).

All three species with evidence of admixture, *E. coriacea*, *E. wachenheimii*, and *E. parviflora*, have been described as hyperdominant—members of a group of 217 tree species that comprise approximately 50% of the tree numbers and biomass of Amazon forests (ter Steege et al., 2013). *Eschweilera coriacea* is the third most abundant tree species across an Amazon-wide network of forest inventory plots, with an estimated census population size of between four and five billion individuals (ter Steege et al., 2020, 2013) and is the only tree species to be considered hyperdominant in both the Amazon basin and Guiana Shield regions (ter Steege et al., 2013).

Is admixture widespread among species of Eschweilera*?*

Given the sizable gaps in available data on hyperdominant species of *Eschweilera*, additional research is clearly needed to reveal the full extent of admixture among them. We found admixture between two species pairs of hyperdominant *Eschweilera* at two different localities, despite sampling 12 or fewer individuals for any species (Table 3-1). Of the three individuals suspected to be hybrids based on morphology, only one showed evidence of admixture, while three other individuals, one originally determined as *E. coriacea* and two as *E. truncata*, were also found to be admixed (Appendix B). This suggests that trees with or without obvious morphological signs of hybridity may have admixed genomes. Data on the phenology of these species is quite limited but indicates that broad overlap in flowering times during the dry season cannot be ruled out based on existing data (Appendix B) and evidence of admixture clearly demonstrates that phenological overlap can occur in the Amazon basin.

Given the current evidence, the large population sizes of these species, their large (Figure 3-5) and frequently overlapping ranges (Mori et al., 2017), and the prevalence of gene tree conflict in our results (Appendix B), we argue that admixture among *E. coriacea*, *E. wachenheimii*, and *E. parviflora* may be extensive and that future efforts are likely reveal further evidence that admixture has played a role in the evolution of these and possibly other species of *Eschweilera*. However, deeper sampling is necessary to determine the extent of admixture and whether additional species admix. The results of several RT tests showed patterns of gene tree conflict suggestive of ancestral evolutionary reticulations, but that failed to meet our criteria for statistical significance (Table 3-2). Future work that implements explicit tests for admixture with more independent loci may provide stronger evidence regarding whether ancient evolutionary reticulations have occurred in *Eschweilera*. Future sampling efforts with a larger geographic focus could also produce quantitative estimates of gene flow among lineages across the Neotropics, and investigate whether entire populations, rather than individuals, bear genomic signatures of admixture.

Biological implications of admixture among dominant tropical lineages

If admixture is widespread, interspecific gene flow may be an important factor in the evolution of the Parvifolia clade and could shape their reproductive biology, local adaptation, and ecological interactions. Hybridization and introgression can have various outcomes including increasing genetic diversity, sharing of adaptive alleles, and either increasing or decreasing the strength of reproductive isolation barriers (Rieseberg and Wendel, 1993). In some cases, a complete breakdown of reproductive isolation barriers can cause “lineage collapse” or “speciation reversal”, resulting in a new lineage with a mosaic genome (Kearns et al., 2018). Alternatively, if hybrids are inviable, prezygotic isolation barriers may evolve (i.e. reinforcement) or there may be little or no lasting population level effects of hybridization. In the case of *Eschweilera*, current evidence suggests that chloroplast capture may be quite common (Huang et al., 2015; O. Vargas et al., *in prep*), indicating that at least some hybrids are capable of backcrossing with their parent species.

Our results show that morphologically defined species largely correspond to distinctive gene pools in our focal species, even in those that admix. The continued genetic cohesion of admixing species could be due to several factors including hybrid inviability or divergent selection acting on suites of traits that differ among these species. Unfortunately, data about the reproductive biology and ecology of the species found to admix are limited. All three species most often occur in non-flooded forests, though *E. coriacea* appears to tolerate flooding more readily than the other two (Mori and Lepsch-Cunha, 1995). *Eschweilera coriacea* frequently reach the canopy while *E. wachenheimii* are typically smaller and occupy the understory. *Eschweilera parviflora* are most often found in the understory, but can also reach the canopy (Mori et al., 2010 onward). All three species differ somewhat in floral morphology (Figure 3-1; Appendix B) and may attract different pollinators, though observations of floral visitors are lacking for these species (Mori et al., 2010 onward). A better understanding of the nuanced ecological differences among these species may help shed light the selective forces that maintain their genetic separation.

A group of taxa that remain largely distinct despite incomplete reproductive barriers is sometimes called a syngameon (Lotsy, 1925; Suarez-Gonzalez et al., 2018). Several of the best-studied examples of syngameons in trees are found within the oaks (*Quercus*), which hybridize prodigiously (Eaton et al., 2015; Hipp et al., 2020), yet largely retain their cohesion as species (Cavender-Bares, 2019; Hardin, 1975; Kremer and Hipp, 2020) and likely facilitate one another's ecological success through introgression (Leroy et al., 2020). Our results suggest that some members of the Parviflora clade including *E. coriacea*, *E. wachenheimii*, and *E. parviflora* could represent a syngameon, which have been hypothesized to be common in tropical trees (Cannon and Lerdau, 2015; Schmitt et al., 2021), but have not often been documented with genomic evidence. Exchanging genes with other species might facilitate local adaptation across the broad ranges of species like *E. coriacea* (Figure 3-5), but further investigation is needed to test for evidence of a relationship between admixture, species abundances, and ecological amplitude.

Population structure

We found evidence of geographic structure within the hyperdominant species *E. coriacea* (Figire 3-3; Appendix B). In *Structure* analyses, runs with the best posterior probability consistently inferred individuals of *E. coriacea* to correspond to two clusters, with individuals assigned varying proportions of the two clusters depending on where the specimen was collected, in a gradient from Panama to Ecuador and Peru to French Guiana and Brazil (Figure 3-3; Appendix B). We also found evidence to suggest population structure in other species in the Parvifolia clade, including *E. truncata* (Figure 3-3), *E. sagotiana* (Appendix B), *E. collina* (Figure 3-4), and *E. pedicellata* (Figure 3-4), though we note our sampling was not designed to make inferences on geographic structure in these species. Phylogeographic structure is expected within broadly distributed Neotropical trees (Dick and Pennington, 2019) and has previously been uncovered in several species (e.g., Dick and Heuertz, 2008; Nazareno et al., 2019).

Implications for the taxonomy of Amazonian trees

While a reassessment of species limits is outside the scope of this work, our results suggest that our focal species can be robustly identified with the methods we employed (Figures 2-2 and 3-3; Appendix B). Despite the occurrence of admixture in some species, most individuals identified as a focal species clustered with other individuals with the same morphological species identification (Appendix B). Our results therefore suggest that admixture has not led to the widespread erosion of species boundaries within the clade and therefore, morphology can be used to reliably distinguish among most co-occurring species of Lecythidaceae. However, cases in which genomic evidence did not match existing determinations suggest that refined taxonomic and genetic studies may be warranted for some species including *E. coriacea* and *E. micrantha* (Appendix B).

Our results show that morphological determinations for specimens collected outside Reserve 1501 more frequently conflicted with genomic evidence than did determinations for specimens from the intensively studied plot (Appendix B). Intraspecific morphological variability, identification errors, as well as admixture may have contributed to this discordance. Many species in the Parvifolia clade, including the three species for which we find evidence of admixture, have similar vegetative characteristics, overlapping phenology, and are broadly distributed across the Neotropics (Figure 3-5; Appendix B; Mori et al., 2010 onward; Mori et al., 2017). Our results suggest that the methodology employed here might be useful for investigating species delimitation in relation to the geography of broadly distributed tropical tree species. To better characterize species boundaries in tropical trees, studies should explicitly investigate morphological characters in conjunction with genomic evidence, including for admixed individuals and/or populations.

Utility of target capture for studying tropical tree populations

There are several methods available to detect evidence of admixture, each with its own benefits and assumptions. The target capture protocol employed here, with probes specifically designed to recover low copy number, genetically variable loci in Lecythidaceae (Vargas et al., 2019), allowed us to investigate evolutionary history using phylogenetic and Bayesian clustering approaches. Our inferences were based on highly variable coding regions, which may be under natural selection. The effect that targeting such regions has on studies of admixture and species delimitation warrants further study. While employing neutral markers or a larger number of loci may have led to different estimates of ancestry, our dataset allowed us to identify admixed individuals and distinguish intraspecific geographic variation from admixture using an explicit test. There are drawbacks to using target capture at infraspecific phylogenetic scales, including the relatively high per sample cost compared to other reduced-representation genome sequencing approaches such as RADseq. However, RADseq protocols often require relatively high-molecular weight input DNA (Graham et al., 2015), while target capture can more readily allow researchers to include samples from partially degraded herbarium specimens (Brewer et al., 2019). We recovered sequences from 343 loci, far fewer than often recovered with RADseq, but far more than most studies that use microsatellites. However, unlike for many RADseq datasets, we recovered sequence data for nearly all target loci for most samples (average 339.6 of 344 loci/individual), which enabled gene-tree based methods for explicitly testing hypotheses of admixture. Our results suggest that target capture can be used to study admixture in topical trees and may be especially useful for studies that wish to include herbarium specimens.

Author Contributions

D.A.L. and C.W.D. conceived the study. A.V. hosted the field work, obtained collection and export permits, and provided access to existing BDFFP collections. D.A.L. and C.W.D. conducted the field work to obtain new collections. D.A.L. conducted the lab work for focal species and designed and performed the analyses. O.M.V. led the sampling and lab work for the broader phylogeny. O.M.V. and D.A.L. mapped collection records. The figures and tables were prepared by D.A.L. The manuscript was written by D.A.L. with editing by C.W.D. and input from all authors.

Acknowledgements

We dedicate this work to the memory of two of our collaborators. Scott A. Mori (1941-2020) was the Lecythidaceae specialist who co-founded the Reserve 1501 Lecythidaceae plot, and kindly and generously supported our work. Paulo A. C. L. Assunção (1956-2021) was a field technician who trained with Scott on the Lecythidaceae plot in the 1980s, became a renowned field botanist in his own right, and trained us in Lecythidaceae field identification. We thank Priscila Souza, Juvenal Batista, Michel Ribeiro, Bruno Garcia Luize, and Xavier Cornejo for contributing collections for our broader sampling. We thank Priscila Souza, Tamara Milton, Nicolli Cabello, and BDFFP Director José Luís Camargo for their help in Brazil and Joseph F. Walker for several helpful discussions about theory. We thank Stephen Smith, Deise Gonçalves, Hannah Marx, Hector Figueroa, Tamara Milton, and three anonymous reviewers for their comments on a previous version of the manuscript. Financial support came from NSF FESD 1338694 to D.A.L. and O.M.V. and NSF DEB 1240869 to C.W.D. Field work by D.A.L. was also supported by the Global Fellowship in Agricultural Development within CA&ES at the University of California, Davis. Samples from Brazil were part of a collaboration (CNPq AEX 01300.006387/2017-42) between the University of Michigan and the National Institute of Amazonian Research (INPA) and were registered in SISGEN-Brazil (#AA0B72D) by A.V. Collecting at BDFFP reserves was conducted under ICMBio permit #58484-1. This is study number 819 of the Technical Series of the Biological Dynamics of Forest Fragments Project (BDFFP - INPA/STRI).

Data Availability Statement

The data, scripts, and output files that support the findings of this study are openly available from Dryad at doi:10.5061/dryad.fj6q573t4. Raw sequence reads are available from NCBI BioProject PRJNA641333.

Figures

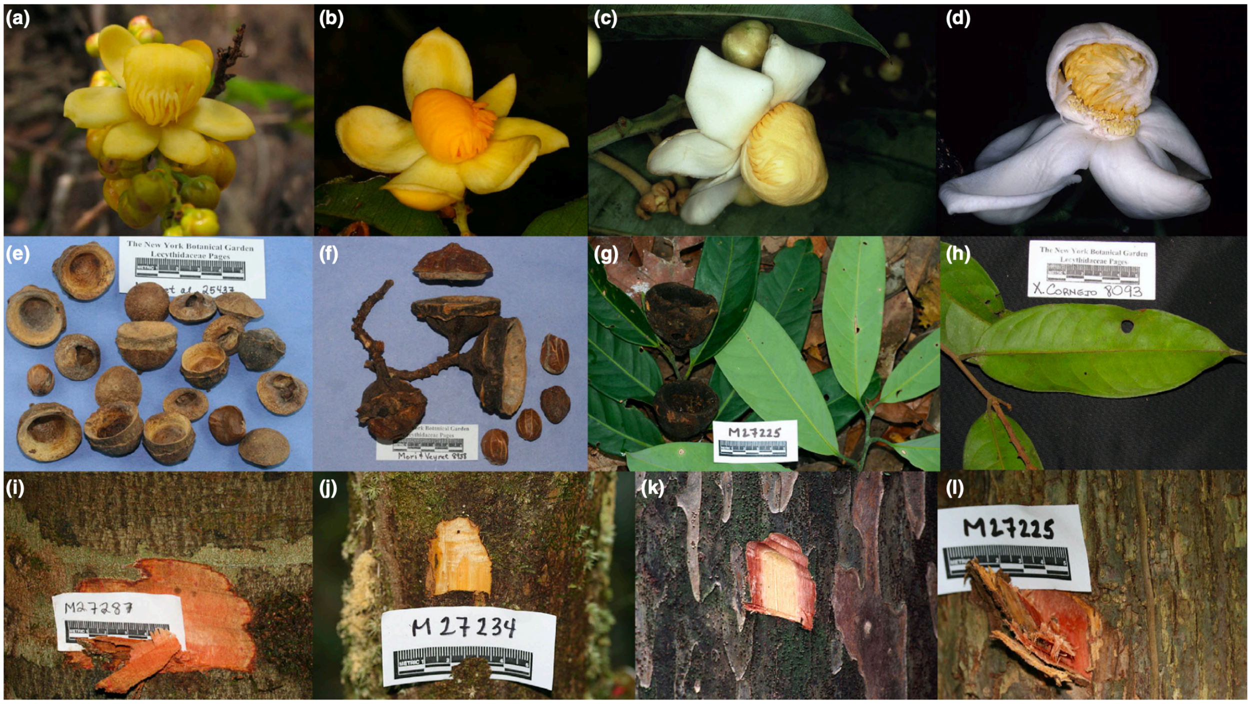
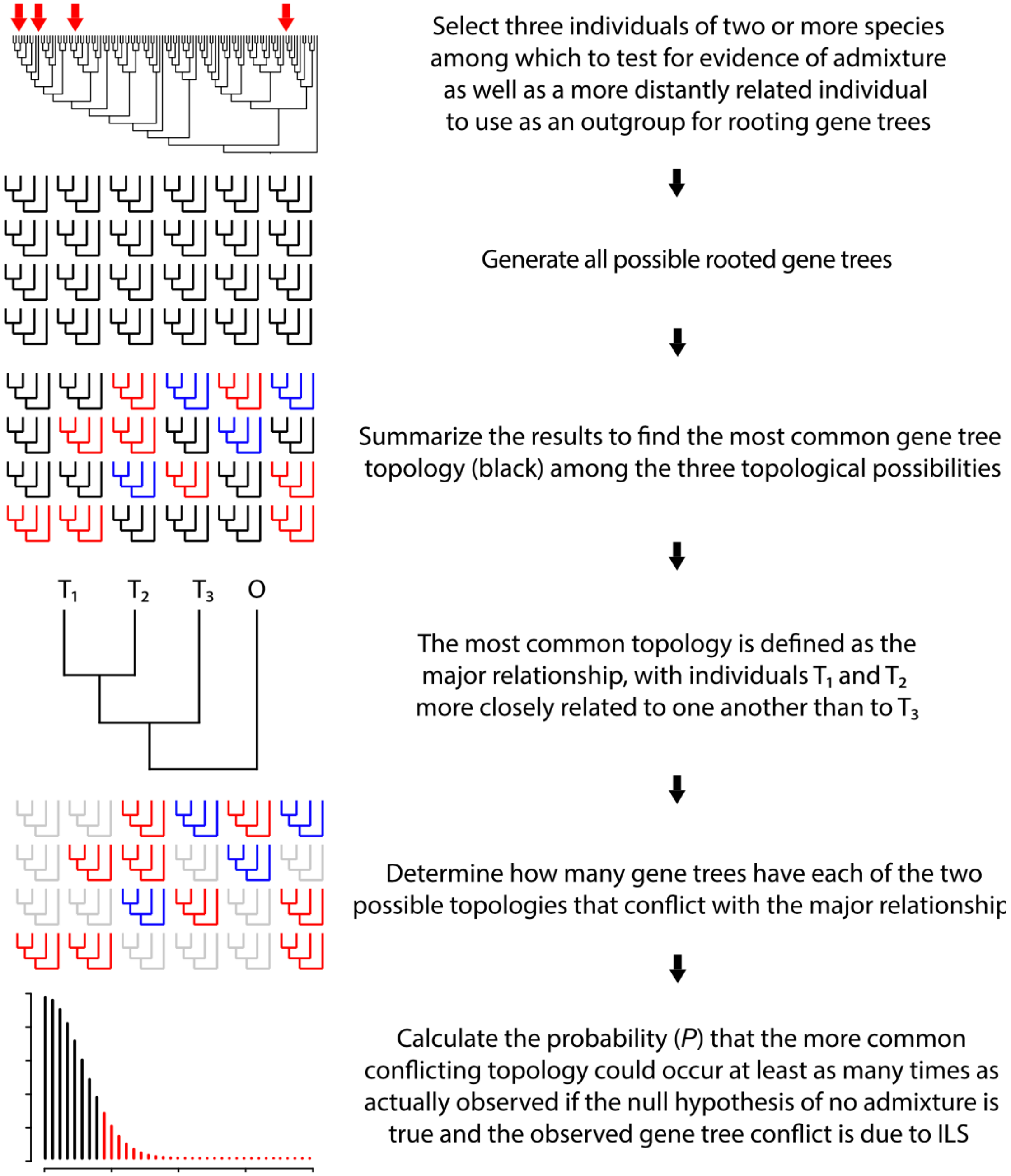


Figure 3-1. Examples of the morphology of members of the Parvifolia clade. A) Flower of *Eschweilera parviflora*. B) Lateral view of a flower of *Eschweilera wachenheimii*. C) Flower of *Eschweilera coriacea*. D) Flower of *Eschweilera collina* with androecial hood sectioned. E) Fruit bases, opercula, and seeds of *Eschweilera parviflora*. F) Fruits, operculum, and seeds of *Eschweilera coriacea*. G) Leaves and old fruit of *Eschweilera atropetiolata*. H) Abaxial view of a leaf of *Eschweilera coriacea*. I) Bark of *Eschweilera tessmannii*. J) Bark of *Eschweilera truncata*. K) Bark of *Eschweilera sagotiana*. L) Bark of *Eschweilera atropetiolata*. Photo attribution: A, B, E, F, G, I, J & L to Scott Alan Mori; C, D & K to Carol Ann Gracie; H to Xavier Cornejo. Reproduced under terms of the CC BY-NC-SA 3.0 license. Captions are adapted from Mori *et al*. (2010 onward).



Select three individuals of two or more species among which to test for evidence of admixture as well as a more distantly related individual to use as an outgroup for rooting gene trees

Generate all possible rooted gene trees

Summarize the results to find the most common gene tree topology (black) among the three topological possibilities

The most common topology is defined as the major relationship, with individuals T_1 and T_2 more closely related to one another than to T_3

Determine how many gene trees have each of the two possible topologies that conflict with the major relationship

Calculate the probability (*P*) that the more common conflicting topology could occur at least as many times as actually observed if the null hypothesis of no admixture is true and the observed gene tree conflict is due to ILS

Figure 3-2. Schematic of the rooted triplet test for assessing evidence of admixture. Red arrows indicate four hypothetical samples selected for the test. The test assumes that the outgroup diverges first in all gene trees and at least two species are represented in the ingroup. Blue and red phylogenies represent the two possible topologies that conflict with the most common topology after all possible gene trees have been generated. Any statistically significant deviation from equal numbers of the two conflicting topologies, where *P* is the probability of a result at least as unequal as the observed frequencies using a binomial test, is considered evidence that the assumptions of the multispecies coalescent have been violated by admixture among species.

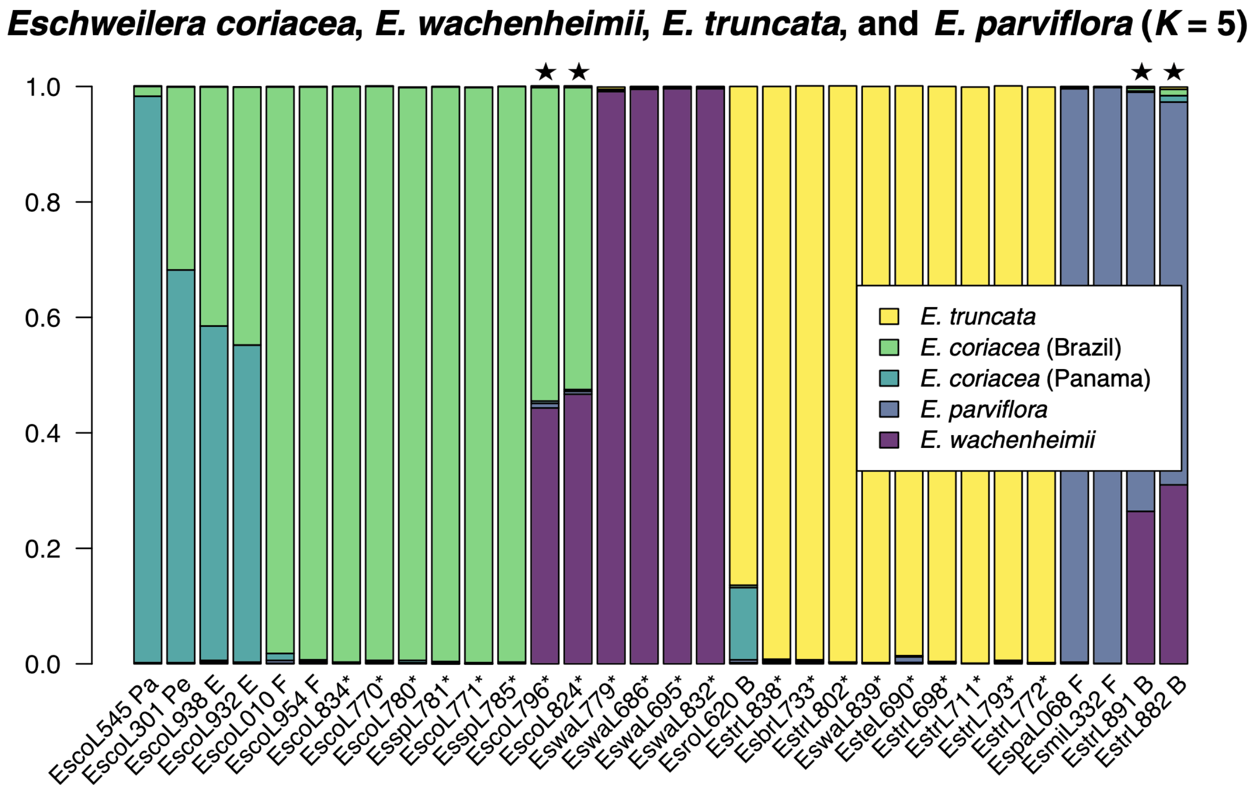


Figure 3-3. Population structure (K=5) of all samples in the clade that included *E. coriacea*, *E. wachenheimii*, *E. truncata*, and *E. parviflora*. Each bar represents the ancestry of an individual inferred with *Structure*. Each individual is labeled with a unique code used throughout all analyses and asterisks indicate samples from focal species collected at Reserve 1501. Collection locations outside Reserve 1501 are indicated as follows: Pa-Panama, Pe-Peru, E-Ecuador, F-French Guiana, B-Brazil. Black stars above bars indicate individuals with significant evidence of admixture based on an RT test.

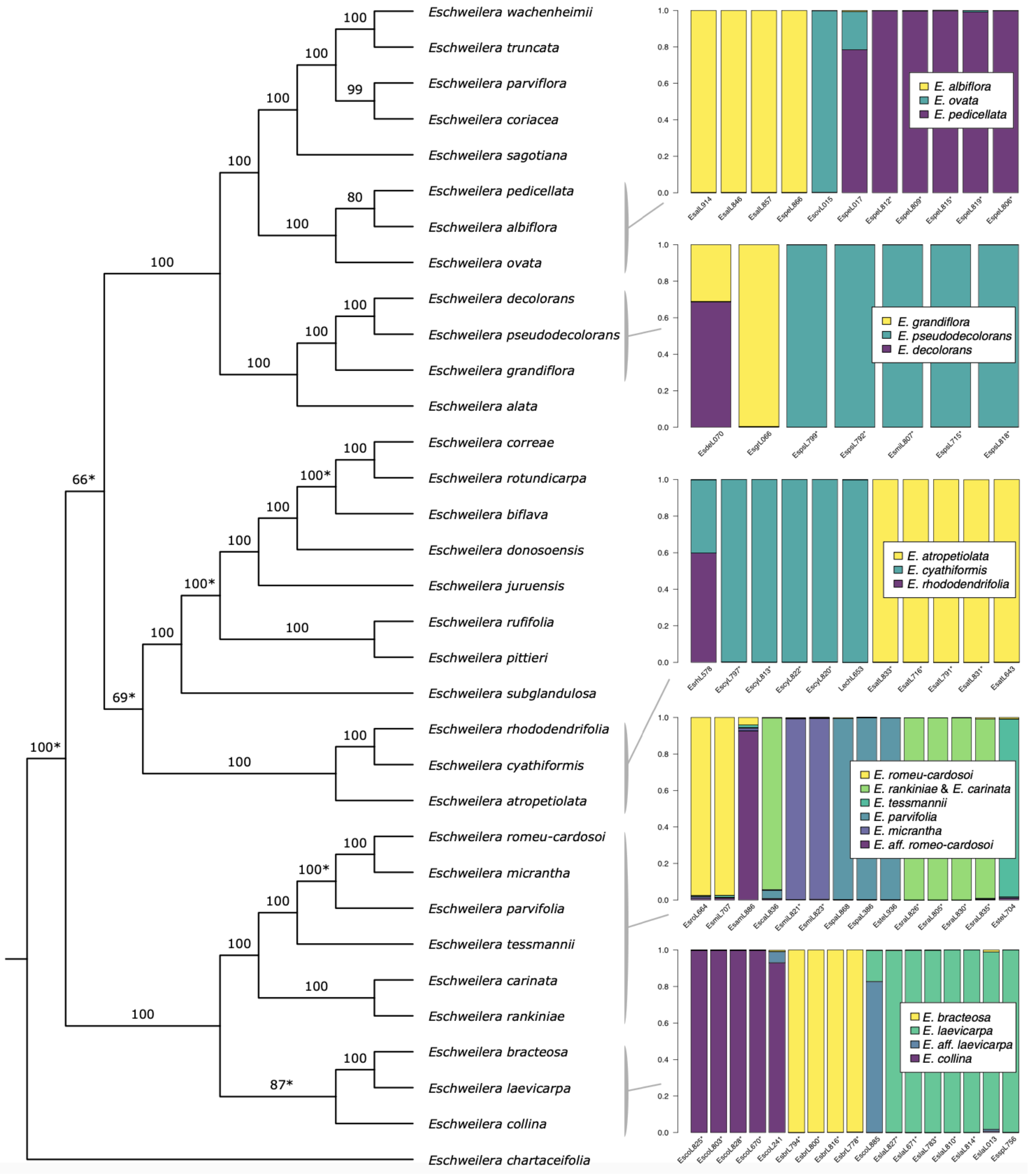


Figure 3-4. Phylogeny of the Parvifolia clade visualized using a single representative per accepted species. Branch labels are IQ-TREE ultrafast bootstrap support. Asterisks on branch labels indicate nodes that conflict with the best scoring maximum likelihood topology recovered with an exon-only supermatrix. The results of *Structure* analyses are shown for SNP datasets that included all individuals within the corresponding clades indicated on the phylogeny. Each individual is labeled with a unique code used throughout all analyses and asterisks on these labels indicate samples from focal species collected at Reserve 1501. The legend for each sub-plot indicates the one or more species that most closely corresponded to each cluster based on accepted taxonomy. The individuals in these clades generally clustered along morphologically defined species boundaries and there was no significant evidence of admixture for these taxa based on rooted triplet tests.

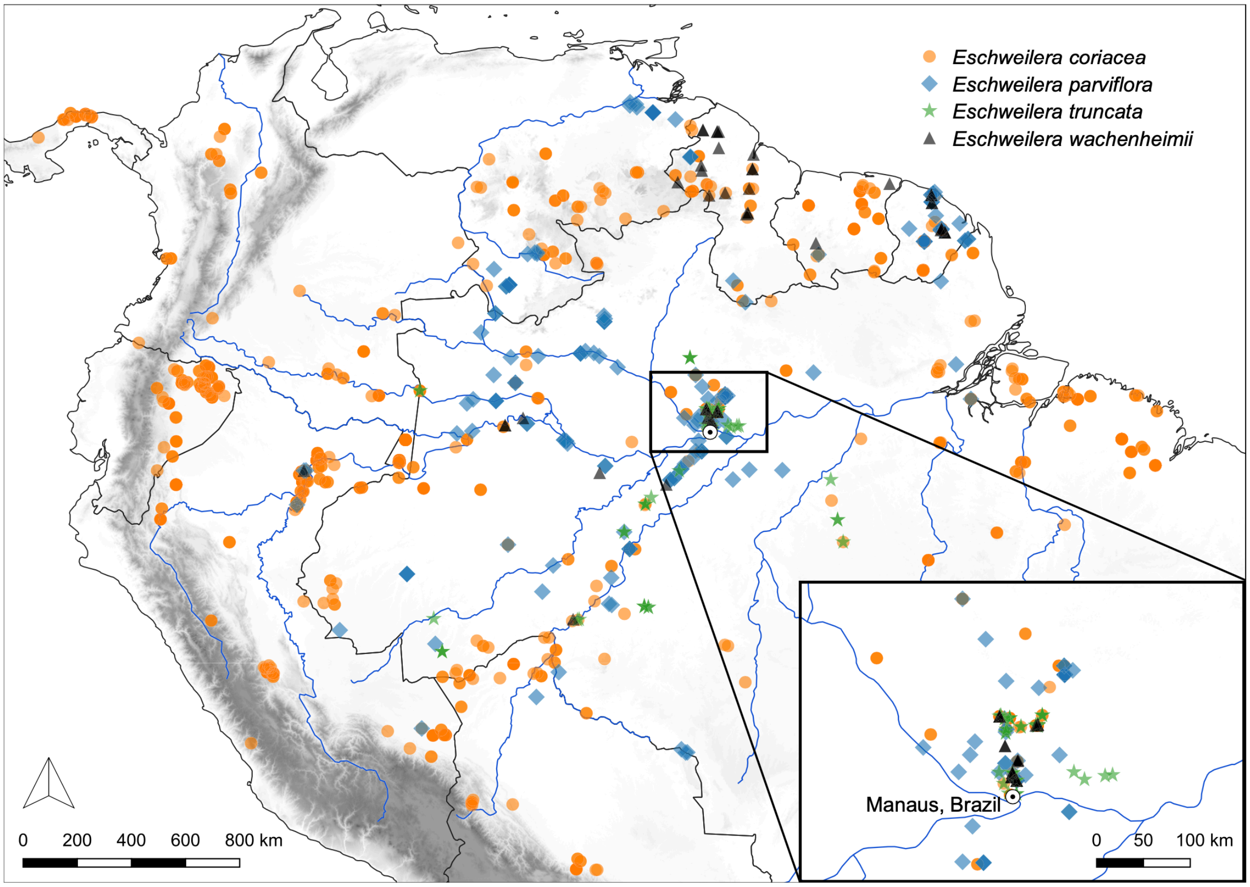


Figure 3-5. Occurrence records across the Neotropics for four closely related species in the Parvifolia clade, three of which show evidence of admixture in this study. Few or no records are available for these species across much of the Amazon basin since most come from collections made in permanent plots. Because of this, there is uncertainty in the true extent of range overlap among these and other species of Lecythidaceae as well as many other clades of Neotropical trees.

Tables

Group or focal species	Number of samples (based on morphology in parentheses)	Named spp. represented
E. atropetiolata S.A.Mori	5 (5)	1
E. bracteosa (Poepp. ex O.Berg) Miers	4 (6)	1
E. collina Eyma	5 (5)	1
E. coriacea (DC.) S.A.Mori	12 (13)	1
E. cyathiformis S.A.Mori	5 (4)	1
E. laevicarpa S.A.Mori	7 (6)	1
E. micrantha (O.Berg) Miers	2 (6)	1
E. pedicellata (Rich.) S.A.Mori	6 (7)	1
E. pseudodecolorans S.A.Mori	5 (4)	1
E. rankiniae S.A.Mori	4 (4)	1
E. truncata A.C.Sm.	10 (9)	1
E. wachenheimii (Benoist) Sandwith	4 (6)	1
Focal species or admixed from Reserve 1501	58 (60)	12
Admixed within Parvifolia clade	4 (3)	n.a.
Parvifolia clade	109 (107)	33
Lecythidaceae	240 (240)	127

Table 3-1. Summary of the number of samples before and after making redeterminations.

Samples forming major relationship	**Third ingroup**	**Major relationship count**	**Conflict 1 count**	**Conflict 2 count**	***P* value**	**Corrected crit. value**	**Reject H-null**
EswaL779, EscoL796	EscoL834	141	124	53	4.87E-08	2.22E-04	Yes
EstrL882, EswaL779	EspaL068	134	124	56	2.18E-07	2.27E-04	Yes
EswaL779, EscoL824	EscoL834	129	126	59	4.69E-07	2.33E-04	Yes
EstrL891, EspaL068	EswaL779	131	113	64	1.42E-04	2.38E-04	Yes
EsmiL332, EspaL068	EswaL779	208	69	40	3.52E-03	2.44E-04	No
EscoL885, EslaL783	EsbrL794	147	101	69	8.58E-03	2.50E-04	No
EstrL838, EscoL834	EssaL335	140	104	73	0.012	2.56E-04	No
EsroL664, EsamL886	EsmiL823	120	111	83	0.026	2.63E-04	No
EscoL241, EscoL828	EscoL885	224	50	36	0.080	2.70E-04	No
EscoL771, EswaL839	EstrL711	212	55	42	0.111	2.78E-04	No
EsteL690, EstrL772	EspaL068	222	50	38	0.120	2.86E-04	No
EstrL838, EswaL779	EssaL335	170	80	66	0.141	2.94E-04	No
EspaL386, EspaL868	EsteL704	221	49	39	0.169	3.03E-04	No
EscyL797, EsrhL578	EsatL643	164	77	65	0.178	3.13E-04	No
EstrL838, EswaL779	EscoL834	118	107	93	0.179	3.23E-04	No

Table 3-2. Summary of 15 rooted triplet tests, ranked in order of increasing *P* value.

Chapter IV

The Phylogeny and Global Biogeography of Primulaceae Based on High-Throughput DNA Sequence Data

Preamble: This chapter has not yet been published elsewhere. The citation for this chapter is: Larson, D.A., Chanderbali, A.S., Maurin O., Goncalves, D.J.P., Dick, C.W., Soltis, D.E., Soltis, P.S., Fritsch P.W., Clarkson, J.J., Grall, A., Davies, N.M.J., Larridon, I., Kikuchi, I.A., Forest, F., Baker W.J., Smith, S.A., Utteridge, T.M.A. 2022. The phylogeny and global biogeography of Primulaceae based on high-throughput DNA sequence data.

Abstract

The angiosperm family Primulaceae is morphologically diverse and is distributed nearly worldwide. However, phylogenetic uncertainty has limited our ability to identify where major morphological and biogeography transitions have occurred. We used target capture sequencing with the Angiosperms353 kit, tree-based sequence curation, and multiple phylogenetic approaches to investigate the major clades of Primulaceae and their relationship to other Ericales. Our sampling included 150 Primulaceae specimens, comprising nearly all recognized genera of the family, with a particular focus on the subfamily Myrsinoideae. We used fossil and secondary calibrations to generate a dated phylogeny and conducted a broad scale biogeographic analysis.

Our analyses resolved relationships among most genera and showed that subfamilies Myrsinoideae and Primuloideae are sister to one another, with Theophrastoideae and Maesoideae successively sister to those. We found unequivocal evidence that *Ardisia*, the largest genus in the family, is non-monophyletic, with at least 19 smaller genera nested within it. *Myrsine*, *Primula*

and *Androsace* are also rendered non-monophyletic by smaller genera. We show that the clade formed by Neotropical *Ardisia* and allies is sister to a group most diverse in the Pacific Islands, suggesting a history of trans-oceanic dispersal. The phylogeny of the family suggests that multiple independent transitions to an herbaceous habit have occurred, or that an early ancestor of Primuloideae and Myrsinoideae was herbaceous.

Our results provide a robust hypothesis for the phylogenetic relationships among the genera of Primulaceae as well as the biogeographic history of its clades. A major taxonomic revision of Myrsinoideae is necessary to make all genera monophyletic. Denser sampling of some genera is necessary to establish whether they are monophyletic as well as whether all *Ardisia* and allied genera that occupy the Neotropics form a single clade.

Introduction

The angiosperm family Primulaceae *sensu lato* (*s.l.*). comprises more than 2600 species, which are distributed nearly worldwide and span a wide variety of ecologies (Stevens, 2001 onward). Morphological diversity within the family ranges from tropical trees (e.g., the rainforest tree *Ardisia copelandii* Mez, which can grow up to 45m), to lianas, woody shrublets, alpine cushion plants, and herbs (Figure 4-1). Species of several genera, including *Primula* L., *Cyclamen* L., and *Androsace* L., are cultivated as ornamentals. Others, including *Embelia* Burm.f*., Myrsine* L*., Ardisia* Sw*.,* and *Lysimachia* Tourn. ex L., have a history of use in traditional medicine (Quattrocchi, 2012). *Aegiceras corniculatum* (L.) Blanco, a member of subfamily Myrsinoideae, is one of only a few angiosperm lineages to have evolved a mangrove habit and is widely distributed along coastal marine ecosystems of Indo-Malaysia, Australia, and the Pacific Islands. Some species of *Ardisia* and *Lysimachia* are considered problematic invasives outside their native ranges (Muñoz and Ackerman, 2011). The genera *Hottonia* Boerh.

ex L. and *Samolus* L. have evolved to live in wet and even aquatic habitats, and some species of *Samolus* are also salt tolerant (Ståhl, 2004). This extensive diversity makes Primulaceae an exceptional system to study evolutionary patterns in morphology, life history, and biogeography.

Our understanding of the evolutionary relationships among the lineages that now comprise Primulaceae and their position in the broader angiosperm phylogeny has changed dramatically in the past three decades with the application of molecular phylogenetic methods. The Cronquist (1981) system circumscribed what is now Primulaceae as its own order, Primulales, which contained three "primuloid families": Primulaceae *sensu stricto* (*s.s.*), Theophrastaceae, and Myrsinaceae. In this system, temperate herbaceous taxa were primarily considered to be members of Primulaceae *s.s.*, while tropical woody taxa, including the Paleotropical *Maesa* Forssk., were included in Myrsinaceae (Cronquist, 1981; Judd et al., 1994). Anderberg and Ståhl (1995) presented a cladistic analysis based on morphology that suggested *Maesa* did not form a monophyletic group with other Myrsinaceae, within which it had usually been regarded as a tribe or subfamily (Mez, 1902). Early phylogenetic analyses also suggested *Maesa* did not form a clade with Myrsinaceae (Anderberg et al., 1998; Morton et al., 1996).

In the first publication by the Angiosperm Phylogeny Group (APG I, 1998), the primuloid families were moved to Ericales *s.l.* because DNA evidence had clarified that they are nested within that clade (Källersjö et al., 1998; Morton et al., 1996). Soon after, Källersjö et al. (2000) used data from three chloroplast genes and found strong support that *Maesa* is sister to the rest of the primuloid clade. Anderberg et al. (2000) then formally recognized Maesaceae as a monogeneric family distinct from Myrsinaceae. Anderberg et al. (2000) also transferred *Samolus* from Primulaceae *s.s.* to Theophrastaceae, and the mostly herbaceous *Lysimachia*, *Cyclamen*, *Coris* L., and *Ardisiandra* Hook.f., as well as others now usually considered synonyms of

Lysimachia (i.e., *Anagallis* L., *Trientalis* L., *Glaux* L., *Asterolinon* Hoffmanns. & Link, and *Pelletiera* A.St.-Hil.), from Primulaceae *s.s.* to Myrsinaceae based on molecular evidence (Manns and Anderberg, 2009). APG II (2003) followed the four-family circumscription of Anderberg et al. (2000), including the treatment of *Samolus* a member of Theophrastaceae, although some authors have continued to regard *Samolus* as its own family (Ståhl, 2010, 2004).

APG III (2009) revisited the classification of the primuloid families and combined the group into a single Primulaceae *s.l.*, in part because DNA evidence had led to such dramatic changes in the historical circumscriptions of Myrsinaceae and Primulaceae *s.s.* (APG III, 2009). Thus, in the APG III (2009) and APG IV (2016) systems, the former families were regarded as the subfamilies Primuloideae, Myrsinoideae, Theophrastoideae, and Maesoideae. The monophyly of this inclusive circumscription has been supported by numerous recent studies (Larson et al., 2020 [Chapter II of this dissertation]; Leebens-Mack et al., 2019; Rose et al., 2018) and is followed here except in historical contexts where noted. Primulaceae have often been inferred to be sister to Ebenaceae, with Sapotaceae sister to these two families (e.g., Rose et al., 2018; Schönenberger et al., 2005). However, a recent analysis using transcriptomic data, including primarily nuclear genes, suggested that support for these relationships was weak and that the position of Primulaceae among the other ericalean families remains uncertain (Larson et al., 2020).

Within Primulaceae, recent studies have suggested that *Ardisia*, the largest genus with ca. 700 species, is not monophyletic (Julius et al., 2021; Rose et al., 2018; Yang and Hu, 2022). These studies have focused primarily on Paleotropical species and have used data from few genes, and thus relationships among Neotropical *Ardisia* and closely related genera remain largely unexplored. There is thus a particular need for investigation that includes broad sampling across the whole subfamily Myrsinoideae. Resolving these relationships will advance understanding of the global biogeographic and morphological history of Primulaceae.

Morphologically based generic concepts have led some authors to suggest that several taxa in Myrsinoideae have disjunct amphi-Pacific tropical distributions, including *Hymenandra* A.DC. ex Spach (Pipoly and Ricketson, 2000, 1999), but whether these represent recent trans-oceanic dispersal events, outcomes of ancient boreotropical forests, or artifacts of taxonomic practice, have not been rigorously assessed. Lack of phylogenetic resolution among members of Myrsinoideae, including *Stimpsonia*, *Coris*, and *Ardisiandra*, has limited our ability to identify transitions in morphology (Wanntorp et al., 2012), such as herbaceous versus woody habit and capsular versus drupaceous fruits. Whether *Oncostemum* A.Juss. is rendered paraphyletic by *Badula* Juss., both of which are endemic to islands of the Indian Ocean, is uncertain. This phylogenetic uncertainty obscures our understanding of the evolution of floral morphology in that clade (Bone et al., 2012; Strijk et al., 2014). Poor understanding of relationships among other morphologically similar taxa with overlapping ranges such as *Tapeinosperma* Hook.f. and *Discocalyx* (A.DC.) Mez have also limited our understanding of the evolution in those groups, including transitions between different mating systems, which have historically been used as diagnostic characters to differentiate genera.

In this study, we generated the first broadly inclusive phylogeny of Primulaceae using high-throughput DNA sequence data to investigate the biogeographic history of the clade. Beyond presenting an updated phylogenetic hypothesis for relationships among nearly all genera of Primulaceae, we seek to answer the following questions: 1) What is the biogeographic history of the pantropical genus *Ardisia* and other morphologically similar woody genera of Myrsinoideae?, 2) What lineages gave rise to various island endemic taxa including *Badula* and

Oncostemum (Indian Ocean Islands), *Pleiomeris* A.DC. and *Heberdenia* Banks ex A.DC. (Canary Islands), and *Tapeinosperma*, *Discocalyx*, *Elingamita* G.T.S.Baylis, and *Loheria* Merr. (Malaysian, Australasian, and Pacific Islands)? and 3) Where have transitions between herbaceous versus woody habit and capsular versus drupaceous fruits occurred since the divergence of Primulaceae from other Ericales? We also discuss the extensive taxonomic implications of our results, particularly highlighting the need for a major generic revision of the subfamily Myrsinoideae.

Materials and Methods

Sampling

We sampled from several herbaria, including the Royal Botanic Gardens, Kew (K), the University of Michigan (MICH), the Florida Museum of Natural History (FLAS), the New York Botanical Garden (NY; Appendix C). In addition, we obtained DNA aliquots for 97 specimens from the Kew DNA Bank. Much of our sampling from K and the Kew DNA Bank included genera from across Ericales and was conducted as part of the Plant and Fungal Tree of Life (PAFTOL) project (Baker et al., 2021). About 30 preserved tissue samples were collected in silica gel, but the majority (ca. 65% of all samples) were from dried herbarium material. We sampled 133 species of Primulaceae, representing 49 of 55 accepted genera (Appexdix C). The six genera we were not able to include comprise 13 accepted species in total (POWO, 2022).

Our final sampling included 324 individuals from Ericales, 150 of which are members of Primulaceae (Appendix C). We sampled 49 accepted genera of Primulaceae and several more previously recognized genera, now usually considered synonyms (e.g., *Dodecatheon* L., *Gentlea* Lundell, *Synardisia* (Mez) Lundell, *Anagallis*, *Glaux*, *Trientalis*; Appendix C). All but six accepted genera (POWO, 2022) within Primulaceae were included in our final phylogenetic analyses. We were not able to sample *Votschia* B.Ståhl (Theophrastoideae, one species, Panama), *Mangenotiella* M.Schmid (Myrsinoideae, one species, New Caledonia), *Solonia* Urb. (Myrsinoideae, one species, Cuba), or *Vegaea* Urb. (Myrsinoideae, one species, Hispaniola). We sequenced a member of *Ctenardisia* Ducke (Myrsinoideae, five species, Neotropics) and *Amblyanthus* A.DC. (Myrsinoideae, four species, Assam, India), but excluded these from our final analyses because of low gene recovery (see below).

Sequencing and assembly

DNA was extracted from tissue with various modified CTAB protocols (Baker et al., 2021). Illumina sequencing libraries enriched for the Angiosperms353 target gene regions were generated with the Angiosperms353 kit (Johnson et al., 2019). Enriched libraries were sequenced with Illumina machines to generate either 150-base pair (bp) or 250-bp paired-end reads. Some samples were sequenced in multiple sequencing runs, and all reads for these samples were concatenated into a single pair of read files per sample prior to analysis.

Reads were trimmed with Trimmomatic (Bolger et al., 2014) by using the options "ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:2:TRUE SLIDINGWINDOW:5:20 LEADING:20 TRAILING:20 MINLEN:36" to trim adapters and bases with a PHRED score of less than 20. Trimmed reads were spot checked with FastQC (Andrews, 2010). Forward and reverse reads that became unpaired because of trimming were combined into a single file per sample. Sequences for target loci were assembled using forward, reverse, and unpaired reads with HybPiper v1.3.1 (Johnson et al., 2016) under default settings and a custom "353mega" target file designed for asterids (McLay et al., 2021). The HybPiper pipeline includes Exonerate (Slater and Birney, 2005), BLAST+ (Camacho et al., 2009), Biopython (Cock et al., 2009), BWA (Li and Durbin, 2009), SAMtools (Li et al., 2009), GNU Parallel (Tange, 2011), and SPAdes (Bankevich et al., 2012). Assembly summary statistics, intron data, and paralog information were generated with python scripts included in HybPiper.

Sequence filtering and paralog removal

We employed a variety of approaches to remove paralogs and sequences that were short or suspected to result from assembly errors (Table 4-1). We used MAFFT v7.310 (Katoh and Standley, 2013) to align non-curated exon data generated with the *retrieve_sequences.py* script in HybPiper and generated a gene tree for each of the 353 genes using RAxML v8.2.12 (Stamatakis, 2014). We then generated a "Preliminary phylogeny" with these gene trees using ASTRAL v5.6.2 (Zhang et al., 2018). Based on the Preliminary phylogeny and inspection of the preliminary gene trees, we decided to include in downstream analyses only samples with at least 10 genes recovered at ≥50% of their expected length. We did this because samples with fewer than 10 such genes were often inferred to have dubious phylogenetic placements in the preliminary results, consistent with lack of phylogenetic signal. We also noted 14 additional samples from across Ericales to exclude from our final analyses because of suspected errors in specimen identification or sample processing errors (Appendix C).

To generate a robust phylogeny for downstream analyses, we used the *paralog_investigator.py* and *paralog_retriever.py* scripts from the HybPiper package to generate a fasta file for each of the 353 genes that included all assembled exon sequences. We removed all copies of sequences flagged as potential paralogs and all sequences that were shorter than 25% of the average length of sequences for that gene using the pxrmt command in phyx (Brown et al., 2017a) and the custom script *remove_flagged_paralogs_and_25pct_short_seqs.py*. Next, the poor-quality samples described above were removed from all gene fasta files using the pxrms function in phyx. Each fasta was then aligned with the E-INS-I algorithm in MAFFT under the options “--maxiterate 1000” and “--genafpair”, and a gene tree was produced for each with IQ-TREE v1.6.12 (Nguyen et al., 2015) and the GTR+Γ model of sequence evolution.

To further reduce the possibility of including paralogs or poorly assembled sequences, we trimmed the resulting exon gene trees using a modified version of the script *trim_trees_based_on_branch_distributions.py* (Larson et al., 2021 [Chapter III of this dissertation]) that uses *trim_tips.py* and *cut_long_internal_branches.py* from Yang and Smith (2014). For each tree, we cut internal branches that were longer than 0.25 substitutions per site or were 10 times longer than their sister branch, keeping only the largest subtree per gene. The 0.25 cutoff was selected based on exploratory analyses of gene trees that showed branches longer than that usually subtended “clades” that were extremely unlikely to reflect a true genetic history (i.e., a long branch subtending a clade nested within a different family) and were therefore likely the result of paralogs or sequence assembly errors. We trimmed terminal branches in gene trees that were longer than three standard deviations above average for terminal branches in that tree using a modified version of the script *trim_trees_based_on_branch_distributions.py* from Larson et al. (2021). Sequences for each remaining tip were used to generate a fasta file for each gene with a slightly modified version of the script *generate_fasta_from_tree.py* from Larson et al. (2021) and were again aligned with MAFFT using the settings described above. A supermatrix for this dataset consisting of the resulting curated exon data for all 353 genes was constructed with the pxcat command in phyx. This dataset is herein referred to as the Ericales1-exon set.

Ericales-wide phylogenetic analyses

The Ericales1-exon supermatrix was used for phylogenetic analyses with IQ-TREE. Three tree searches were conducted, each with a separate model partition for each of the 353 genes, with Edge-linked-equal, Edge-linked-proportional, and Edge-unlinked partition models specified with the "-q", "-spp", and "-sp" flags, respectively (Chernomor et al., 2016). We also conducted an ultrafast bootstrap analysis for all maximum likelihood (ML) trees generated in this study and consider 95% as the threshold for strong support (Minh et al., 2013). Gene trees for each exon alignment were also estimated separately with IQ-TREE and used to generate a species tree with ASTRAL.

Primulaceae-specific dataset construction and phylogenetic analyses

To investigate phylogenetic relationships within Primulaceae more thoroughly, and test the effect of including data from introns, we generated two additional datasets using the 150 Primulaceae samples that met our threshold for gene recovery with two Ebenaceae and two Sapotaceae samples comprising the outgroup. We retrieved intron data for all samples using the *intronerate.py* script included with HybPiper. We used a modified version of the script *parvifolia_alignment_from_tree.py* (Larson et al., 2021) to subset the intron sequences so as to only include those with a corresponding exon sequence in the Ericales1-exon set. That is, if the exon sequence for a particular gene was removed for a particular sample during construction of the Ericales1-exon dataset described above, the corresponding intron data for that gene were not included in further analyses. Both the exon and intron sequences were subset with the pxrms command in phyx so as to only include the 150 Primulaceae in the Ericales1-exon dataset and the four samples of the outgroup. Intron and exon sequences for each gene were aligned separately with the E-INS-i algorithm in MAFFT. Then, columns with >50% missing data were removed from all intron alignments using the pxclsq command in phyx because introns had a high proportion of missing data, which is expected because of the nature of target capture sequencing. We refer to the resulting exon-only dataset as the Prim2-exon set and the dataset that included both intron and exon data as the Prim3-intron set.

The Prim2-exon dataset contained only the exon sequences from 150 samples of Primulaceae and the four samples of the outgroup and was therefore an exact subset of the Ericales1-exon dataset, but with sequences realigned with MAFFT. The Prim3-intron set included the same realigned exon data as well as corresponding intron data. Genes trees for the Prim2-exon set were generated with IQ-TREE and the GTR+Γ model. Gene trees for the Prim3-intron set were generated by concatenating the intron and exon sequences for each gene and estimating a phylogeny in IQ-TREE with separate GTR+Γ model partitions for the intron and exon data (i.e., two partitions per gene). A species tree was estimated for both sets of gene trees with ASTRAL. A supermatrix for each dataset was constructed by concatenating alignments for each gene. For each supermatrix, three tree searches were conducted with IQ-TREE, each with a separate model partition for each of the 353 exon and intron sequences (where applicable) with the Edge-linked-equal, Edge-linked-proportional, and Edge-unlinked partition models specified with the “-q”, “-spp”, and “-sp” flags, respectively.

Selection of preferred model

We generated four phylogenies for each of our three datasets, one using ASTRAL and three using maximum likelihood methods with different partitioning models implemented in IQ-TREE. For all three datasets (Ericales1-exon, Prim2-exon, and Prim3-intron), the Edge-linked-proportional model received the best Bayesian Information Criterion (BIC) score, with the Edge-linked-equal model second best. The Edge-unlinked run received the best Akaike Information Criterion (AIC) score and the best corrected AIC score for all datasets, seemingly because AIC and corrected AIC do not penalize additional parameters to the same extent as BIC. The Edge-linked-proportional partition model includes one additional parameter for each model partition relative to the Edge-linked-equal partition model. This extra parameter scales the branch lengths to account for heterotachy but keeps branch lengths proportional to one another (Chernomor et al., 2016). The Edge-unlinked model instead includes an entirely separate set of branch lengths for each model partition, which dramatically increases the number of parameters. For example, for the Ericales1-exon dataset with 353 model partitions, the number of free parameters was 3822, 4174, and 188,770 for the Edge-linked-equal, Edge-linked-proportional, and Edge-unliked partition models, respectively. For that dataset, that averaged to approximately 535 parameters per supermatrix partition (i.e., per gene) for the Edge-unliked model as compared to 11 or 12 for the Edge-linked-equal and Edge-linked-proportional models, respectively.

Robinson-Foulds (RF) distances among all pairs of trees were calculated with the bipartition analysis program *bp* from the gophy package (Robinson and Foulds, 1981; https://github.com/FePhyFoFum/gophy/). For visualization, a phylogeny of Ericales with one sample per family was produced with the Ericales1-exon-ML and the pxrmt command in phyx. The RF distances showed that the topologies recovered with the Edge-linked-equal and Edge-linked-proportional models were identical for the Prim2-exon and Prim3-intron datasets and were very similar for the Ericales1-exon dataset. Given this and the BIC scores, we chose the Edge-linked-proportional model as our preferred model for all datasets and herein limit our discussion of supermatrix analyses to these results, referring to these phylogenies as the maximum likelihood tree (ML) for each dataset (i.e., the Ericales1-exon-ML, Prim2-exon-ML, Prim3-inton-ML trees).

Time-calibrated phylogeny

We generated an ultrametric tree with branch lengths corresponding to time using a penalized likelihood method implemented in the program treePL (Sanderson, 2002; Smith and O'Meara, 2012). We used the maximum likelihood phylogeny for the Ericales1-exon dataset (i.e., Ericales1-exon-ML) for this analysis because its branch lengths are proportional to molecular substitutions and its topology was qualitatively similar to those of most other trees generated (Appendix C). Prior to running treePL, we trimmed the Ericales1-exon-ML tree to include only samples from Primulaceae and the four samples of the outgroup described above and trimmed duplicate samples from the same species.

Although the fossil record for Primulaceae is somewhat sparse (Boucher et al., 2016), there are fossils that can be used for time-calibration. Friis et al. (2010) described a primuloid flower mesofossil found near the present-day municipality of Mira, Portugal, which the available evidence suggests is from Campanian–Maastrichtian (83.6–66.0 Ma) sediments of the Late Cretaceous (Walker et al., 2013). Boucher et al. (2016) used this fossil to set a maximum age of 72 million years ago (Ma) for the clade of Primulaceae that includes *Androsace*, *Primula*, and *Soldanella* L. This 72-Ma age corresponds to the approximate time of the boundary between the Campanian and Maastrichtian (Walker et al., 2013). Friis et al. (2021) later formally described this primuloid fossil as *Miranthus elegans* E.M.Friis, P.R.Crane & K.R.Pedersen within Primulaceae *s.l.* Friis et al. (2021) also described a second species of *Miranthus* from the same locality and noted that *Miranthus* bears particular similarity to the extant genus *Samolus*. This similarity conceivably warrants the placement of *Miranthus* at a node within the crown Primulaceae (e.g. Theophrastoideae); however, because Friis et al. (2021) found that there are other placements that are nearly as parsimonious and did not assign the fossils to an extant genus, we chose to place *Miranthus* at the crown node of Primulaceae (Table 4-2).

Boucher et al. (2016) included several additional calibrations based on fossilized seeds. Those authors assigned *Androsace* a minimum age of 5.3 Ma based on seeds from the Miocene and *Primula* a minimum age of 15.97 Ma based on seeds from *Primula rosiae* from the middle Miocene. Because of sampling idiosyncrasies, they assigned the group formed by *Androsace* + *Primula* + *Soldanella* a minimum age of 28 Ma based on fossil seeds of *Lysimachia angulata* from the Early Oligocene. Rose et al. (2018) included broader sampling in their analysis, and instead placed the *L. angulata* fossil at the crown node of the clade that includes *Lysimachia* and *Trientalis* (often considered a synonym of *Lysimachia*), bounding the age to a minimum and maximum of 28.1 and 33.9 Ma, respectively. Rose et al. (2018) placed the *P. rosiae* fossil at the crown node of the clade of *Primula* that included *P. meadia* (L.) A.R.Mast & Reveal and *P. sieboldii* É.Morren, bounding the minimum and maximum age of that clade to between 11.6 and 16.0 Ma, respectively (i.e., the middle Miocene). Because our sampling also includes *Lysimachia*, we followed Rose et al. (2018) and assigned the *L. angulata* fossil to the crown of the *Lysimachia s.l.* clade, except that we only assigned a minimum age of 28.1 Ma. We assigned the *P. rosiae* fossil to the crown node of the *Primula* clade in our study and assigned a minimum age of 16.0 Ma. For the minimum age of crown *Androsace*, we followed Boucher et al. (2016) and assigned a minimum age of 5.3 Ma.

If no point calibration is used, molecular dating approaches based on penalized likelihood can suffer from unidentifiability, where multiple sets of model parameter values fit the data equally well and with no way to distinguish among them (Yang, 2014). There is no known Primulaceae fossil that can be used as a reliable point calibration because its age and precise placement on the phylogeny would need to be known without error. Because an ideal point calibration was not possible, we used a secondary calibration for the stem age of the family based on the results of Magallón (2015). The stem age of Primulaceae is the time at which the family began evolving as an independent lineage after it diverged from its sister lineage, which Magallón et al. (2015) found to be Ebenaceae. That study used both penalized likelihood and Bayesian methods to produce a time-calibrated phylogeny of angiosperms and found that the estimated stem age of Primulaceae varied with the method used. Penalized likelihood resulted in a point estimate of 94.92 Ma and a confidence interval of between 91.92 and 98.88 Ma, based on dated bootstrap trees. Using a Bayesian method, they recovered 86.95 Ma as the median estimate for the stem age, with minimum and maximum ages of 71.64 and 97.45 Ma, respectively, based on the 95% highest posterior probability distribution. We chose to use the point estimate of 94.92 Ma from the penalized likelihood analysis of Magallon et al. (2015) as a point calibration for the stem age of Primulaceae in our study. As compared to the results of their Bayesian analysis, this date is more in line with the results of Rose and colleagues, who estimated this age to be approximately 100.9 Ma (Rose et al., 2018; Figure 3 of that study).

We estimated uncertainty in our divergence times by generating bootstrapped trees with IQ-TREE, the same settings described above, and with the topology fixed to that of the Ericales1-exon-ML tree. We conducted two bootstrap analyses with branch length optimization, each with 100 replicates. The first was a standard non-parametric bootstrap analysis with the "-b" option, and the second was a standard non-parametric bootstrap analysis that resampled both partitions and sites within partitions with the "-bsam GENESITE -wbtl" options. We dated all bootstrapped trees with treePL and the settings and calibrations described above and summarized the results with TreeAnnotator from the BEAST2 package (Bouckaert et al., 2014). Uncertainty reported here refers only the 95% highest density interval (HDI) from the second analysis (which resulted in larger uncertainty intervals then the first analysis).

Biogeographic dataset construction and analysis

We conducted a biogeographic analysis using the DEC* model (Massana et al., 2015; Ree and Smith, 2008) implemented in BioGeoBEARS (Matzke, 2013) and based on the time-calibrated phylogeny described above. We used the biogeographic realms of Olson et al. (2001). Our regions were 1) Nearctic, 2) Palearctic, 3) Neotropic, 4) Afrotropic, 5) Indo-Malay, 6) Oceania, and 7) Australasia. Our biogeographic analysis was therefore similar to that of Rose et al. (2018) except that we only included Primulaceae in our analysis, used a different variation of the DEC model, and used different boundaries for some realms, including Oceania as distinct from Australasia. We primarily based our scoring on the native ranges of species as described in the Plants of the World Online Database (POWO, 2022). We also considered additional information such as regional online floras and the Global Biodiversity Information Facility (www.gbif.org), including for species that occurred near transition zones between realms.

We generally excluded a species from a realm when only a small part of its range crossed a boundary. This situation occurred near the Nearctic/Neotropic boundaries in North America and the Palearctic/Indo-Malay boundary in east Asia. As an example, a species with a Caribbean distribution that also extended into northern Florida, USA, would be considered only Neotropic, rather than Nearctic as well.

Results

Gene recovery

We recovered 105,115 assembled exon sequences using HybPiper, for an average of 297.8 sequences per each of the 353 target genes (including sequences flagged as paralogs). After removing paralogs and low-quality samples, we retained 95,646 exon sequences (271.0 per gene on average). After phylogeny-based trimming, we retained 93,326 sequences (264.4 per gene on average). Full occupancy for exons would be 324 samples each with 353 sequences or 114,372 total sequences; therefore, our Ericales1-exon dataset had an average gene occupancy of 83.6%. Recovery for introns was more limited in that for the 150 Primulaceae and four samples of the outgroup in the Prim3-intron dataset, we recovered 17,577 intron sequences (an average of 49.8 sequences per gene) out of the 54,362 possible (32.3% occupancy). The average length of intron alignments was 627.1 bp (Appendix C).

Phylogenetic relationships

When not otherwise noted, phylogenetic relationships discussed refer to those recovered in the Ericales1-exon-ML tree (i.e. our focal tree). Relationships among all families of Ericales were supported by 100% ultrafast bootstrap support, except for a sister relationship between Primulaceae + Ebenaceae and Polemoniaceae + Fouquieriaceae, which received 84% ultrafast bootstrap support (considered poor support; Figure 4-2). Within Primulaceae, Myrsinoideae was sister to Primuloideae, and Maesoideae was sister to the rest of the family (Figure 4-3).

Relationships within subfamily Myrsinoideae—The pantropical *Ardisia* was highly non-monophyletic in our results, such that 19 other accepted genera were interdigitated among its members (Figures 4-3 and 4-4). Because of this and other relationships, it is difficult to apply existing taxonomic names. For example, the tribe Ardiseae as defined by Mez (1902), whose work still represents the most recent monographic treatment of Myrsinoideae as a whole, includes *Aegiceras*, *Ardisia*, *Heberdenia*, *Conandrium* (K.Schum.) Mez, and *Hymenandra*. Mez's tribe Myrsineae contained most other woody Myrsinoideae, including *Badula*, *Loheria*, *Myrsine*, and *Tapeinosperma*. Both tribes are non-monophyletic in our results. Therefore, for clarity, we assign informal names to certain strongly supported clades of Myrsinoideae. We use the term "Ardisioids" to refer to the clade that includes *Ardisia* and *Tapeinosperma* (Figure 4-3). Within this clade, we refer to the "New World Ardisioids" as the clade that includes *Ardisia tinifolia* Sw. and *Stylogyne* A.DC. We refer to the clade that includes *Discocalyx* and *Badula* as the "Old World Ardisioids" (Figure 4-3). The Old World Ardisioids excludes *Elingamita* and *Tapeinosperma*, because they are more closely related to the New World Ardisioids. In addition, we use the terms "Myrsinoids" to refer to the clade that includes *Myrsine*, *Cybianthus* Mart., and *Embelia*. This usage of "Myrsinoids" is not equivalent to subfamily Myrsinoideae but is instead more restrictive, nor is it synonymous with most uses of tribe Myrsineae (Figure 4-3). For the more inclusive clade that includes *Aegiceras, Monoporus* A.DC., and all *Ardisia* we use the term "Woody Myrsinoideae" (Figure 4-3).

In all ML and ASTRAL trees, *Elingamita* and *Tapeinosperma* were sister to one another and formed a clade with the New World Ardisioids. *Loheria* and *Discocalyx* formed a clade that was sister to the other Old World Ardisioids in five of the six trees, but not in Ericales1-exon-ASTRAL where *Loheria* was instead sister to all other Ardisioids. *Aegiceras* and the Madagascar endemic *Monoporus* formed a clade that was sister to all other members of the Woody Myrsinoideae in five of six trees. The exception was in Prim2-exon-ASTRAL where *Monoporus* was instead sister to the clade formed by *Elingamita*, *Tapeinosperma* and the New World Ardisioids, but was subtended by a branch that was near-zero in length and had only 42% support (i.e., ASTRAL local posterior probability). *Geissanthus* Hook.f., *Parathesis* Hook.f., *Stylogyne*, and *Wallenia* Sw. were inferred to belong to the New World Ardisioids. We sampled one species of *Hymenandra*, the Neotropical *Hymenandra pittieri* (Mez) Pipoly & Ricketson, which was also nested within the New World Ardisioids. *Amblyanthopsis* Mez, *Antistrophe* A.DC., *Badula*, *Conandrium*, *Discocalyx*, *Emblemantha* B.C.Stone, *Fittingia*, *Labisia* Lindl., *Loheria* (except in Ericales1-exon-ASTRAL)*,* *Oncostemum*, *Sadiria* Mez, and *Systellantha* B.C.Stone were within the Old World Ardisioids. The phylogeny of the Ardisioids displayed a clear geographic signal, and the sampled members of *Ardisia* were clearly differentiated as members of either the New World Ardisioids or Old World Ardisioids (Figures 4-3 and 4-4).

Lysimachia and *Cyclamen* were sister to each other in all six trees. *Ardisiandra* was sister to *Lysimachia* + *Cyclamen* in four trees (87% ultrafast bootstrap support in our focal tree), but not in Prim2-exon-ML or Prim3-intron-ASTRAL, in which *Ardisiandra*, *Coris*, and *Stimpsonia* were successively sister to all other Myrsinoideae. *Stimpsonia* was sister to all other Myrsinoideae in all trees, except in Prim2-exon-ASTRAL, where it was sister to Primuloideae, but with only 67% local posterior probability, which is generally considered poor support (Sayyari and Mirarab, 2016).

Relationships within Theophrastoideae and *Primuloideae*—Within Theophrastoideae, tribe Theophrasteae Bartl. comprises all genera except *Samolus* (i.e., *Bonellia* Bertero ex Colla*,* *Deherainia* Decne., *Jacquinia* L.*,* *Theophrasta* L.*,* *Neomezia* Votsch, *Clavija* Ruiz & Pav., and the unsampled genus *Votschia*). Sampled members of tribe Theophrasteae formed a monophyletic group with two subclades, each with three genera. *Bonellia* and *Deherainia* were collectively sister to *Jacquinia*, whereas *Clavija* was sister to *Theophrasta* + *Neomezia*. Four trees, including our focal tree, placed *Samolus* sister to tribe Theophrasteae, whereas the Ericales1-exon-ASTRAL and Prim2-exon-ASTRAL trees instead placed *Samolus* sister to Primuloideae + Myrsinoideae.

Within Primuloideae, *Dionysia* Fenzl and *Kaufmannia* Regel were nested within *Primula* in all six trees. The topology within the subfamily was identical among all trees, except for one node within the clade formed by *Primula, Kaufmannia,* and *Dionysia* (Figure 4-3). The monospecific *Pomatosace* Maxim. was nested within *Androsace* in all trees, and *Androsace* + *Pomatosace* was sister to the rest of Primuloideae. *Omphalogramma* Franch. and *Bryocarpum* Hook.f. & Thomson were sister to each other, with *Soldanella* and *Hottonia* successively sister to those.

Diversification times and biogeographic analyses

The crown age of Primulaceae was inferred to be 77.1 Ma (Figure 4-4). The crown ages of Maesoideae, Theophrastoideae, Primuloideae, and Myrsinoideae were estimated to be 10.2 Ma, 65.6 Ma, 55.0 Ma, and 56.6 Ma, respectively (Table 4-3). The crown age of the Ardisioids was inferred to be 24.7 Ma, with the New World Ardisioids and Old World Ardisioids dating to 20.0 Ma and 23.5 Ma respectively (Figure 4-4).

The ancestral area of Myrsinoideae was inferred to be the Palearctic, with *Lysimachia* colonizing the Neararctic several times by approximately 41.3 Ma. The ancestral area of the Woody Myrsinoideae was inferred to be Indo-Malaysia, whereas the MRCA of the Ardisioids and that of the Myrsinoids were both inferred to have occupied the Indo-Malay and Neotropical realms, implying two separate dispersal events to the Neotropics ca. 28.7 and 24.7 Ma. The Old World Ardisioids were inferred to have begun diversifying in the Indo-Malay realm, with various lineages later dispersing to the Oceania, Australasia, Afrotropic, and Palearctic realms. The MRCA of *Tapeinosperma, Elingimeta*, and the New World Ardisioids was inferred to have occupied the Neotropic and Australasian realms.

The ancestral area of *Maesa* was inferred to be the Palearctic and Indo-Malay realms, with subsequent expansion into the Afrotropics. The ancestral areas of Mysinoideae + Primuloideae and of Myrsinoideae + Primuloideae + Theophrastoideae were inferred to be the Palearctic. Within Primuloideae, *Primula* and *Hottonia* were inferred to have arrived in the Neararctic by 13.7 and 14.9 Ma, respectively. The ancestor of *Samolus* and Theophrasteae was inferred to have been Neotropical. The MRCA of *Samolus* was also inferred to have been Neotropical, with the genus then diversifying and expanding to its current, near-cosmopolitan distribution.

The ancestral area of Primulaceae was inferred to have been both the Palearctic and Neotropics; however, there were several states with similar marginal probabilities for this node. The marginal probability of the most likely state, Palearctic and Neotropic, was 0.0540554 followed by Palearctic only (0.0512223), Palearctic, Indo-Malay, and Neotropic (0.03439358), Palearctic, Nearctic, and Neotropic (0.03043411), and Palearctic, Neotropic, and Afrotropic (0.02902059). For comparison, the state with the 10th highest marginal probability for the root of Primulaceae was Palearctic, Nearctic, Neotropic, and Afrotropic (0.02120134).

Discussion

Phylogeny of Primulaceae

We performed the first phylogenetic analysis based on high-throughput DNA sequencing designed to address phylogenetic and biogeographic questions across Primulaceae. We found a different topology for relationships among Primulaceae, Ebenaceae, Sapotaceae, Polemoniaceae + Fouquieriaceae, and Lecythidaceae than several recent studies (Larson et al., 2020; Leebens-Mack et al., 2019; Stull et al., 2020; Zhang et al., 2020). While our results offer a new hypothesis for the relationship of Primulaceae relative to other Ericales, it remains one hypothesis among many, reflecting the extensive phylogenomic conflict that has been shown to characterize family-level divergences in the order (Larson et al., 2020). Our results agree with previous studies regarding the relationships among the currently accepted subfamilies of Primulaceae (e.g., Källersjö et al., 2000; Larson et al., 2020; Leebens-Mack et al., 2019; Rose et al., 2018). However, in our results, support for a sister relationship between *Samolus* and tribe Theophrasteae was not definitive.

The phylogenetic placement and taxonomic treatment of *Samolus* has historically been somewhat controversial. *Samolus* and species sampled from the tribe Theophrasteae were sister in our focal tree and received 95% ultrafast bootstrap support, which is the minimum threshold for significance at a 5% false positivity rate (Minh et al., 2013). Two of our six trees (Ericales1-exon-ASTRAL and Prim3-exon-ASTRAL) placed *Samolus* sister to Primuloideae + Myrsinoideae rather than sister to tribe Theophrasteae. A morphological cladistic analysis by Anderberg and Ståhl (1995) placed *Samolus* as a lineage sister to most other Primulaceae *s.s.* (now subfamily Primuloideae). Early molecular phylogenetic analyses suggested that *Samolus* did not form a clade with Primulaceae *s.s.* but was instead likely sister to Theophrastaceae (Källersjö et al., 2000). Caris and Smets (2004) conducted a detailed morphological and developmental analysis of *Samolus* and concluded that there are no “unambiguous morphological characters” that are synapomorphies shared between *Samolus* and tribe Theophrasteae. A possible synapomorphy could be the presence of staminodes, which occur in both, but may develop through different ontogenetic pathways in the two groups (Caris and Smets, 2004). Furthermore, staminodes appear to be evolutionarily labile in Primulaceae because they are absent in some species of *Samolus* (Wanntorp et al., 2012), and are also present in *Soldanella* (Primuloideae) and sometimes in *Maesa* (Anderberg et al., 1998; Anderberg and Ståhl, 1995; Friis et al., 2021). Ståhl (2004, 2010) treated *Samolus* as a distinct family in his taxonomic treatments of Theophrastaceae. Future studies should continue to investigate the strength of the evidence for a sister relationship between Theophrasteae and *Samolus*.

Our sampling includes all genera within subfamily Theophrastoideae except for the monospecific, Panama endemic, *Votschia*, which is clearly a member of tribe Theophrasteae based on its morphology (Ståhl, 1993). The other five genera of Primulaceae not included here are well supported as members of Myrsinoideae based on morphological evidence (Schmid, 2011; Ståhl and Anderberg, 2004). We do not consider the recently proposed *Paralysimachia* F.Du, J.Wang & S.S.Yang as an accepted genus of Primulaceae (Du et al., 2016). This is because phylogenetic evidence is lacking and the single proposed species exhibits morphology atypical for the family, including that it has basal placentation and no corolla (versus free-central placentation as a synapomorphy in Primulaceae and corolla present in most species; Judd et al., 2015; Schönenberger et al., 2005). Neither do we recognize *Evotrochis* Raf., which was recently proposed by transferring *Primula* subgenus *Sphondylia* Duby to its own genus because of the group’s close affinity to *Dionysia* (Firat and Lidén, 2021). Our results and other studies (e.g., Rose et al., 2018) have shown that *Dionysia* is nested within *Primula*. Therefore, unless a future revision of *Primula* more clearly justifies the recognition of *Evotrochis* separate genus, these species can remain as *Primula* subgenus *Sphondylia*.

We consider the relationship of *Stimpsonia* as sister to all other Myrsinoideae to be strongly supported by our results, although in one tree (i.e., Prim2-exon-ASTRAL), the genus was recovered as sister to Primuloideae with weak support. The placement of *Ardisiandra* was somewhat uncertain; it was weakly supported as sister to *Cyclamen* + *Lysimachia* in our focal tree (87% ultrafast bootstrap support), as well as in three of five other trees, but not in Prim2-exon-ML or Prim3-intron-ASTRAL, where it was sister to all other Myrsinoideae except *Coris* and *Stimpsonia*.

Our results suggest that *Pleiomeris* and *Heberdenia*, two monospecific genera endemic to the Canary Islands, are nested within *Myrsine*. Two exceptions in our results were in Ericales1-exon-ASTRAL where *Pleiomeris* + *Heberdenia* was sister to all sampled *Myrsine* and in Prim2-exon-ASTRAL where *Heberdenia* was sister to *Myrsine* + *Pleomeris*. However, all ML trees recovered the two as nested within *Myrsine* with 100% ultrafast bootstrap support. Appelhans et al. (2020) also found that both *Pleiomeris* and *Heberdenia* were nested within *Myrsine,* as did Yang and Hu (2022). Julius et al. (2021) found that *Pleiomeris* was nested within *Myrsine* using ITS but did not sample *Heberdenia*. Both genera have also been placed into, and moved out of, *Myrsine* at various times in their taxonomic history. The clearest taxonomic solution is to include them as members of *Myrsine*.

Relationships within the Ardisioids and their systematic implications

A particular strength of our sampling was for *Ardisia* and its close relatives that together formed the "Ardisioids." We sampled nearly all accepted genera of Myrsinoideae as well as the type species and several subgenera of *Ardisia* (POWO, 2022). Our results unequivocally demonstrate that *Ardisia* as currently circumscribed is not monophyletic, nor are the tribes Ardiseae and Myrsineae of Mez (1902). The non-monophyly of *Ardisia* has been suggested from other recent investigations of the genus and other genera based on ITS sequences and several other gene regions (Julius et al., 2021; Yang and Hu, 2022) and an Ericales-wide biogeographic study with publicly available sequences (Rose et al., 2018). We found that 19 other genera were interdigitated with *Ardisia*, which together comprise a strongly supported clade (Figure 4-3). Because they share morphological similarities, we believe several of the five genera of Myrsinoideae that we did not sample including, *Amblyanthus*, *Ctenardisia*, *Mangenotiella*, and *Vegaea*, are likely nested within the Ardisioids as well (Schmid, 2011; Ståhl and Anderberg, 2004).

There were also strongly supported clades within the Ardisioids. All sampled members of *Ardisia* subgenus *Tinopsis* (i.e., *Ardisia sumatrana* Miq., *Ar. purpurea* Reinw. ex Blume, and *Ar. celebica* Scheff.), which is widespread across southeast Asia, formed a clade in all analyses which was sister to the New Guinea endemic *Conandrium*. All samples of *Ardisia* subgenus *Crispardisia* (*Ar. crenata* Sims and *Ar. Polysticta* Miq.) formed a clade, as did samples of subgenus *Icacorea* (*Ar. guianensis* (Aubl.) Mez and *Ar. subsessilifolia* Lundell). The Neotropical subgenus *Ardisia* (*Ar. bracteosa* A.DC., *Ar. revoluta* Kunth, and *Ar. tinifolia*) formed a clade within which *Wallenia* was nested. The Paleotropical species *Ar.* crenata and *Ar. polysticta* formed a clade with *Amblyanthopsis* in all trees.

Julius et al. (2021) sampled 60 mostly Old World taxa of *Ardisia* for their ITS phylogeny and found that many clades they recovered roughly corresponded to the recognized subgenera of *Ardisia*, although there were some exceptions, including cases where other genera were nested within subgenera and where phylogenetic support for monophyletic subgenera was lacking. Yang and Hu (2022) conducted an even larger study of Old World *Ardisia* and their bacterial symbionts and found support for the monophyly of the *Ardisia* subgenera *Crispardisia*, *Pimelandra*, and *Stylardisia* and their newly circumscribed *Bladhia s.str*. and *Odontophylla* (resulting from the transfer of some members of subgenus *Bladhia s.l.* to the new *Odontophylla*). Yang and Hu (2022) also found that the Old World members of *Hymenandra*, *Badula*, *Oncostemum*, and *Sadiria* were nested within *Ardisia* and referred to that clade as the "*Ardisia* generic complex." Our diverse sampling of Ardisioids shows that this "*Ardisia* generic complex" contains more currently recognized genera than has ever been shown before, because it includes all Neotropical members of the Ardisioids as well.

An extensive taxonomic revision of generic limits within the Ardisioids is required. Such a revision is currently underway, including by some co-authors of the present work, with an initial focus on Malaysian taxa (Drinkell and Utteridge, 2015; Dubéarnès et al., 2015; Julius et al., 2021; Julius and Utteridge, 2012; Yang and Hu, 2022). To circumscribe monophyletic genera, one option would be to combine *Ardisia* with most other genera of Myrsinoideae. However, as we note above, there are many well-supported clades within the Ardisioids, many of which display clear morphological and anatomical synapomorphies as well as signals of geographic endemism (Yang and Hu, 2022). The fact that many currently recognized subgenera of *Ardisia* seem to correspond to clades in phylogenetic analyses and have distinct morphological characters that are useful for identification should be taken into consideration in the decision whether to treat all Ardisioids as a single megadiverse, globally distributed genus that contains well over one third of all Primulaceae species.

The alternative is to circumscribe the Ardisioids as several monophyletic genera, though additional studies with very extensive sampling are required before this can be confidently undertaken. This will be an immense task, because *Ardisia* alone currently has over 700 accepted species (POWO, 2022). The Neotropical, Jamaican endemic *Ardisia tinifolia* is the type species of the genus; thus, dividing the Ardisioids into multiple genera would almost certainly result in a newly circumscribed *Ardisia* being restricted to the Neotropics. Many of the architectural and floral characteristics historically used to describe genera within the Ardisioids appear to be homoplasious when viewed in the light of molecular phylogenies (Yang and Hu, 2022). Detailed morphological analysis will likely be necessary in tandem with broadly inclusive molecular studies of the group to identify which characters are synapomorphies for distinct clades and which represent evolutionary convergences. Combining phylogenomic datasets with broader sampling of markers like ITS and chloroplast genes will likely be important for resolving the relationships among the nearly 1000 species of the Ardisioids.

Our single sample each of *Labisia* and monospecific *Emblemantha* were sister to each other in all trees. These genera both superficially appear to be herbaceous (Figure 4-1) but are actually soboliferous shrublets with woody rhizomes or rhizome-like branches that run underground and produce short, leafy shoots. Our results suggest that this habit is a synapomorphy for the clade formed by the two genera. Additional sampling is needed to determine whether *Emblemantha* is nested within *Labisia* (9 spp.; POWO, 2022).

Badula and *Oncostemum*, endemic to islands of the Indian Ocean, were each monophyletic and sister to each other in all analyses. This contrasts with results by Bone et al. (2012) and Strijk (2014), who found that *Badula* rendered *Oncostemum* paraphyletic. Our samples of these genera were from the same specimens as those of Bone et al. (2012) and Strijk

(2014), including for *Oncostemum pachybotrys* Mez and *O. palmiforme* H.Perrier. Their results suggested that *O. palmiforme* was more closely related to *Badula* than to *O. pachybotrys*, whereas we found that both *Oncostemum* and *Badula* were strongly supported as monophyletic. Rose et al. (2018) recovered a nearly monophyletic *Oncostemum* except for a single sample that was sister to *Badula* and was subtended by a near-zero-length branch. Future studies should investigate the cause(s) of this discordance, including whether hybridization and/or introgression might have caused conflicting phylogenetic signal between nuclear and chloroplast genomes and whether this discordance varies across individuals. Alternatively, these results could be due to incomplete lineage sorting resulting from a rapid radiation, or lack of sufficient phylogenetic information in earlier studies. Additional sampling is needed to determine whether all ca. 100 species of *Oncostemum* form a monophyletic group.

Elingamita consists of a single species, *Elingamita johnsonii* G.T.S.Baylis, which occurs in northern New Zealand, and which was recovered as sister to our single sample of *Tapeinosperma* in all analyses. Baylis (1951) described *Elingamita* as a new genus of Myrsinaceae based on floral morphology and considered it to differ from all other genera of the family as circumscribed by Mez (1902). Our sampling precludes us from addressing the question of whether *Elingamita* renders the more diverse *Tapeinosperma* paraphyletic, but the range of extant *Tapeinosperma* demonstrates a clear ability of the genus to disperse among Pacific Islands. Future studies should test whether or not *Elingamita* diverged before the diversification of extant *Tapeinosperma*, especially since, based on our results, the history of these lineages has implications for our understanding of the biogeographic history of the New World Ardisioids. A similar question remains for the monospecific genus *Mangenotiella* from New Caledonia (Schmid, 2011), for which no publicly available DNA sequence currently exists. Phylogenetic evidence is needed to test whether *Mangenotiella stellata* M.Schmid is nested within another described genus, especially *Tapeinosperma*.

Tapeinosperma and *Discocalyx* have often been considered to be closely related based on shared morphology but can be differentiated by bisexual versus unisexual flowers, respectively (Ståhl and Anderberg, 2004). We found that *Tapeinosperma* is more closely related to the New World Ardisioids than to *Discocalyx*, which indicates that either their overall morphologies have converged, or that they have both retained similar ancestral characters of the Ardisioids. *Elingamita* (sister to *Tapeinosperma* in our results) has unisexual flowers and is considered by some to be functionally dioecious (Heenan, 2000), whereas *Tapeinosperma* has bisexual flowers. *Loheria* (most likely sister to *Discocalyx* based on our results) is dioecious, whereas *Discocalyx* is dioecious or bisexual—both genera have unisexual flowers (Ståhl and Anderberg, 2004). Additional sampling of these genera will provide insight into the evolution of mating systems in these groups and other Ardisioids.

Diversification times and biogeographical insights

Our estimated dates for divergence times generally agree with those in other recent analyses (reviewed by Stevens, 2001 onward). We inferred that the MRCA of extant Primulaceae occurred 77.1 Ma, during the Campanian age of the Late Cretaceous and most likely occupied both the Palearctic and Neotropics, but with a marginal probability of only 5.41%. Rose et al. (2018) also inferred that the MRCA of Primulaceae occupied both the Palearctic and Neotropics but with a probability of 43% using a similar model. The fossil genus *Miranthus* is known from the Palearctic (Friis et al., 2021), which is consistent with the inferred ancestral area of Primulaceae. Rose et al. (2018) inferred the ancestral area of Primuloideae +

Theophrastoideae + Myrsinoideae to be the Palearctic and Neotropics as well, whereas in our results, the most likely area for that node was Palearctic only. Our results agree with those of Rose et al. (2018) that the MRCA of *Lysimachia* and other early ancestors of Myrsinoideae occupied the Palearctic. Our biogeographic results regarding Primuloideae also agree with Rose et al. (2018), who found that the subfamily originated in the Palearctic and began colonizing the Nearctic about 16 Ma. Our results agree with previous suggestions that *Samolus* originated in the Neotropics, as did Theophrastoideae (Rose et al., 2018; Stevens, 2001 onward). We inferred a relatively young crown age for *Maesa* (Table 4-3), probably because we had relatively few samples from members of this genus (Appendix C).

The estimated uncertainty intervals in divergence times were generally small, reflecting similar branch lengths (i.e., low variance) for comparable branches among the replicate dated bootstrap trees. This low variance is likely due to the large size of the underlying Ericales1-exon supermatrix (403,832 aligned sites). Bootstrap analyses are expected to converge on high support for values (i.e., the correct values if the model is correctly specified) as the size of the dataset increases (Seo, 2008). Therefore, it is important to recognize that our results suggest that, given the data and the evolutionary model, there is little ambiguity in the data regarding divergence times, and we therefore have *precise* dating estimates. Whether these divergence times are *accurate* in an absolute sense cannot be address with the present approach. Changing the taxon sampling, evolutionary model, and time calibrations would almost certainly lead to different divergence time estimates, which could fall outside the bounds of the 95% HDI we observed. We therefore urge a nuanced interpretation of these estimates since they reflect strong phylogenetic signal in the data, given the model, not necessarily the true uncertainty that exists in these divergence times.

Biogeographic history of the Ardisioids—We found that the New World Ardisioids were sister to a clade formed by *Elingamita* and *Tapeinosperma. Elingamita* is a monospecific, New Zealand endemic, while the more diverse *Tapeinosperma* is native to Malaysia and the southwest Pacific (e.g., Borneo, Fiji, Vanuatu, New Caledonia). The clade of Old World Ardisioids (which does not include *Elingamita* or *Tapeinosperma*) was inferred to have a MRCA that was Indo-Malaysian, from which various lineages later dispersed to Australasia, Oceania, and the Afrotropics. Rose et al. (2018) were not able to clearly resolve the biogeographic history of the Woody Myrsinoideae because of a lack of phylogenetic support, but suggested that some Neotropical *Ardisia* might form a clade, within which some Old World *Ardisia* (including *Ar. glauca* Mez and *Ar. speciosa* Blume) and some other genera might be nested. We found that no Old World *Ardisia* were nested within the New World Ardisioids, which could be due to better phylogenetic support in our study or having not sampled those particular species. Our results agree with Rose et al. (2018) that the lineage that would become *Badula* and *Oncostemum* likely dispersed from Indo-Malaysia to the Afrotropics about 9.8 Ma, although we inferred this date to be slightly older at around 12.3 Ma.

The MRCA of the Ardisioids and Myrsinoids occupied Indo-Malaysia in our results. After divergence from the Myrsinoids, the Ardisioids dispersed to the Neotropics, such that the MRCA of all Ardisioids occupied both the Indo-Malay and Neotropic realms. The Indo-Malaysian population then began diversifying into the Old World Ardisioids, while the Neotropical population dispersed to Australasia, such that the MRCA of the New World Ardisioids and *Elingamita* + *Tapeinosperma* occurred in both the Neotropics and Australasia. The Neotropical lineage then diversified into the New World Ardisioids, and the Australasian population diversified into *Elingamita* and *Tapeinosperma*, the center of diversity of which is now in New Caledonia and nearby Pacific Islands.

We suspect that this model-based inference, that an Ardisioid lineage dispersed from Indo-Malaysia to the Neotropics and then back to Australasia, could be an artifact due to not having sampled any extant Indo-Malaysian *Tapeinosperma*, of which there are at least two species in Borneo (POWO, 2022). Sampling additional Australasian *Tapeinosperma* could also have changed the inferred ancestral range of the Ardisioids, given that we sampled only one of about 79 accepted species of that genus (POWO, 2022). Additionally, our biogeographic analysis did not include distance-based model components, which may have better accounted for the fact that Australasia is directly adjacent to the Indo-Malay realm while the Neotropics were at the time, and still are, several thousand miles away.

A perhaps more parsimonious biogeographic hypothesis is that after the divergence of the lineage that became the Old World Ardisioids, an ancestor of the New World Ardisioids and *Elingamita* + *Tapeinosperma* occupied Indo-Malaysia, from which members of this clade dispersed to Australasia and/or islands in the Pacific and from there dispersed to the Neotropics. Alternatively, dispersal may have occurred directly from Indo-Malaysia to the Neotropics, and separately to Australasia, after divergence from the Old World Ardisioids. Better sampling within *Tapeinosperma* and the Old World Ardisioids is needed to investigate which, if any, scenario is better supported by the phylogeny of the genus and the distributions of extant species.

Disjunct distribution of Hymenandra—Pipoly and Ricketson (1999) transferred nine species of Neotropical *Ardisia* to the previously exclusively Paleotropical genus *Hymenandra* based on stamen morphology and plant architecture. This circumscription meant that *Hymenandra* appeared to have a disjunct distribution between Indo-Malaysia (eight species; Assam, Myanmar, Bangladesh, south-central China, Borneo, and Malaya) and the Neotropics (eight species; Colombia, Nicaragua, Costa Rica, and Panama; POWO, 2022). We found that two samples of *Hymenandra pittieri* were nested within the New World Ardisioids. Julius et al. (2021) sampled two species from Borneo, *Hymenandra rosea* B.C.Stone and *H. beamanii* B.C.Stone, which in their study were nested within a large clade consisting of *Ardisia* subgenera *Pyrgus*, *Tinus*, *Crispardisia*, and *Bladhia* as well the genus *Sadiria*. We show that *Ardisia crenata* (a member of *Ardisia* subgenus *Crispardisia*) and *Sadiria* are nested within the Old World Ardisioids. Yang and Hu (2022) also found an Old World species of *Hymenandra* to be nested within *Ardisia* and pointed out that some morphological characters used to distinguish between the two genera, including the degree to which the anther filaments are fused to one another and the corolla, have evolved multiple times in Myrsinoideae. The degree to which stamens appear to be fused in various members of the Ardisioids may also be influenced by whether the floral material is dried, preserved in alcohol, or observed fresh (T.M.A. Utteridge, personal observation). Based on an analysis of ITS by Julius et al. (2021), the Old World *Hymenandra* may themselves not be monophyletic. In light of this evidence, our results suggest that Neotropical and Paleotropical *Hymenandra* are not closely related and that their “disjunct distribution” is due to the group being non-monophyletic as currently circumscribed.

Implications for the evolution of habit in Primulaceae

There have been multiple transitions hypothesized between woody and herbaceous habits within the history of Primulaceae (Anderberg et al., 1995). Early circumscriptions of family limits used habit to separate the woody Myrsinaceae from herbaceous Primulaceae *s.s.*, but this trait now appears to be highly homoplasious (Källersjö et al., 2000; Table 4-4). The MRCA of all Primulaceae was probably woody, because the family is usually inferred to be sister to Ebenaceae (trees) or else placed along the backbone of Ericales (e.g., Larson et al., 2020), among which most families are woody. *Maesa* is strongly supported as sister to the rest of Primulaceae and are woody shrubs, trees, scrambling climbers, or lianas (Anderberg et al., 2000; Sumanon et al., 2021; Utteridge, 2012).

Members of subfamily Primuloideae are perennial herbs (although some species are somewhat woody at their base) and are sister to Myrsinoideae in our results and previous analyses (Källersjö et al., 2000; Rose et al., 2018). Various lineages of Myrsinoideae have differing habits. Woody members of the Myrsinoideae are mostly shrubs and trees, although some, like the genus *Embelia*, are climbers. As noted above, *Labisia* and *Emblemantha* are woody shrublets with the appearance of herbs, as are some *Ardisia* such as *Ar. primulifolia* Gardner & Champ. Some other lineages of Myrsinoideae are herbaceous.

We found that *Stimpsonia*, which consists of annual herbs, are sister to the rest of Myrsinoideae. *Coris* has been considered either a woody subshrub (Lens et al., 2005; Stevens, 2001 onward) or a perennial herb with a woody base (Judd et al., 2015; Ståhl and Anderberg, 2004) and appears to have diverged from other Myrsinoideae along the backbone of the subfamily soon after *Stimpsonia*. Lens et al. (2005) suggested that the wood of *Coris* is anatomically paedomorphic, such that it appears to have evolved from an herbaceous ancestor, a condition which they term “secondary woodiness.” *Coris* comprises a single species, *Coris monspeliensis* L., which we infer to have diverged from other Myrsinoideae approximately 56.6. Ma. Considering its position in the phylogeny, it may never be possible to determine whether an ancestor of *Coris* was “fully” herbaceous, or if the paedomorphic wood characteristic of the species could have evolved directly from a more typical woody ancestor. For the purposes of our discussion, we consider *Coris* to be herbaceous to emphasize that an anatomic transition appears to have occurred in this species, but we note that it does exhibit a somewhat woody habit (Källersjö et al., 2000; Lens et al., 2005).

Cyclamen (perennial herbs) was sister to *Lysimachia* in all six trees, and this sister relationship received 100% ultrafast bootstrap support in our focal tree. *Lysimachia* species are mostly herbs, though the Hawaiian *Lysimachia* (a clade sometimes referred to as subgenus *Lysimachiopsis*) share a woody shrub habit as a synapomorphy and appears to be deeply nested within otherwise herbaceous clades (Hao et al., 2004). This strongly implies that these woody *Lysimachia* evolved from an herbaceous ancestor, a fact also supported by an analysis of their wood anatomy (Lens et al., 2005).

In most of our results, *Ardisiandra* (herbs) was sister to *Cyclamen* + *Lysimachia* (mostly herbs), a scenario that, assuming woodiness is ancestral, implies that there have been at least three transitions to an herbaceous habit within Myrsinoideae (the others being in *Stimpsonia* and *Coris*). In two of six trees, *Ardisiandra* was inferred to be sister to the Woody Myrsinoideae + *Cyclamen* + *Lysimachia*, a scenario that implies four transitions: one in *Stimpsonia*, a second in *Ardisiandra*, a third in *Coris*, and a fourth in *Cyclamen* + *Lysimachia*. Although our results suggest that they do not form a clade, all herbaceous Myrsinaceae also share capsular fruits, whereas no woody members of the clade exhibit this trait and instead have drupaceous fruits.

It is also possible that the MRCA of Primuloideae + Myrsinoideae was herbaceous, with the woody habit having (re-)evolved in the clade of Woody Myrsinoideae along with drupaceous fruits. *Samolus* species are herbs, with one species having evolved a subshrub habit (Ståhl, 2004). In our focal tree and three others, *Samolus* was sister to a clade corresponding to tribe

Theophrasteae (woody shrubs and trees), which implies that an herbaceous habit evolved after *Samolus* diverged from other Primulaceae. However, as noted above, our results do not entirely exclude the possibility that *Samolus* is instead sister to Primuloideae + Myrsinoideae. If that were the case, it could lend strength to the notion that the common ancestor of Primuloideae, Myrsinoideae, and *Samolus* was in fact herbaceous. Lens et al. (2005) did not find anatomical evidence in species of the Woody Myrsinoideae that they thought suggested herbaceous ancestry. Lens et al. (2005) did note, however, that the wood of Theophrastoideae (recognized as Theophrastaceae in that study) is characterized by very short vessel elements and fibers similar to those found in *Coris* and *Lysimachia* (both of which they thought evolved from herbaceous ancestors), as well as *Aegiceras* (mangroves). Perhaps future anatomical and/or gene expression studies can shed light on the mechanisms responsible for the apparent lability of woodiness and repeated parallel evolution of herbaceousness in Myrsinoideae or offer additional insight into the possibility that a common ancestor of some or all subfamilies of Primulaceae was herbaceous.

Conclusions

Our phylogenomic analyses resolved the major clades of Primulaceae, including the relationships of nearly all genera. Phylogenomic evidence largely supports previous hypotheses of the evolutionary relationships among subfamilies and genera, but with better resolution. This refined phylogenetic hypothesis of Primulaceae highlights numerous taxonomic issues in the group and suggests several directions for future research into the evolution of morphology and mating systems in the clade. The pantropical *Ardisia* is not monophyletic and forms a clade with other genera that arose in Indo-Malaysia approximately 24.7 Ma. Neotropical members of *Ardisia* and several smaller genera form a clade, the MRCA of which arrived in the Neotropics and began diversifying about 20.0 Ma. This Neotropical clade is most closely related to *Tapeinosperma* and *Elingamita*, whose centers of diversity are in islands of the Pacific. *Discocalyx* and *Loheria* are likely closely related and sister to a clade formed by Old World species of *Ardisia* and several smaller genera. An ancestor of the clade formed by the monophyletic genera *Badula* and *Oncostemum* arrived at islands of the Afrotropics from Indo-Malaysia about 12.3 Ma. The Canary Island endemics *Pleiomeris* and *Heberdenia* are closely related, shared a common ancestor about 14.1 Ma, and are nested within the more diverse and widespread *Myrsine*. There have either been parallel transitions to an herbaceous habit in *Samolus*, Primuloideae, and at least three lineages of Myrsinoideae, or a common ancestor early in the history of Primulaceae was herbaceous, with woodiness evolving in the woody clade of Myrsinoideae.

Author contributions

D.A.L. and T.M.A.U. conceived the study. D.A.L., T.M.A.U., P.W.F., D.J.P.G., C.W.D., D.E.S., P.S.S., S.A.S., A.G., and N.M.J.D. conducted the sampling. J.J.C. provided essential lab training to D.A.L. D.A.L. and A.S.C. performed the laboratory work with contributions from J.J.C. and I.A.K. W.J.B., F.F., and O.M. facilitated and supervised the data generation through PAFTOL. S.A.S., A.S.C., C.W.D., D.E.S., P.S.S., and I.L. facilitated and supervised the other data generation. D.A.L. designed and conducted the analyses, wrote the manuscript with input from all authors, and prepared the tables and figures.

Acknowledgements

We thank K, MICH, FLAS, NY, and the Kew DNA bank for access to their collections: without the long-term efforts of countless collectors and curatorial staff, this study would not have been possible. We thank Alexandre Zuntini, Robyn Cowan, Vanessa Barber, and other members of the PAFTOL team for their technical and logistical help. We thank Jacob Berv and members of the Smith and Dick labs for comments on an earlier version of the manuscript. This work was funded in part by grants from the Calleva Foundation to the Plant and Fungal Tree of Life Project (PAFTOL) at the Royal Botanic Gardens, Kew. Support also came from NSF DEB 1917146, DEB 1458466, and DEB 1916555 and from a fellowship to D.A.L. from the Department of Ecology and Evolutionary Biology at the University of Michigan.

Figures

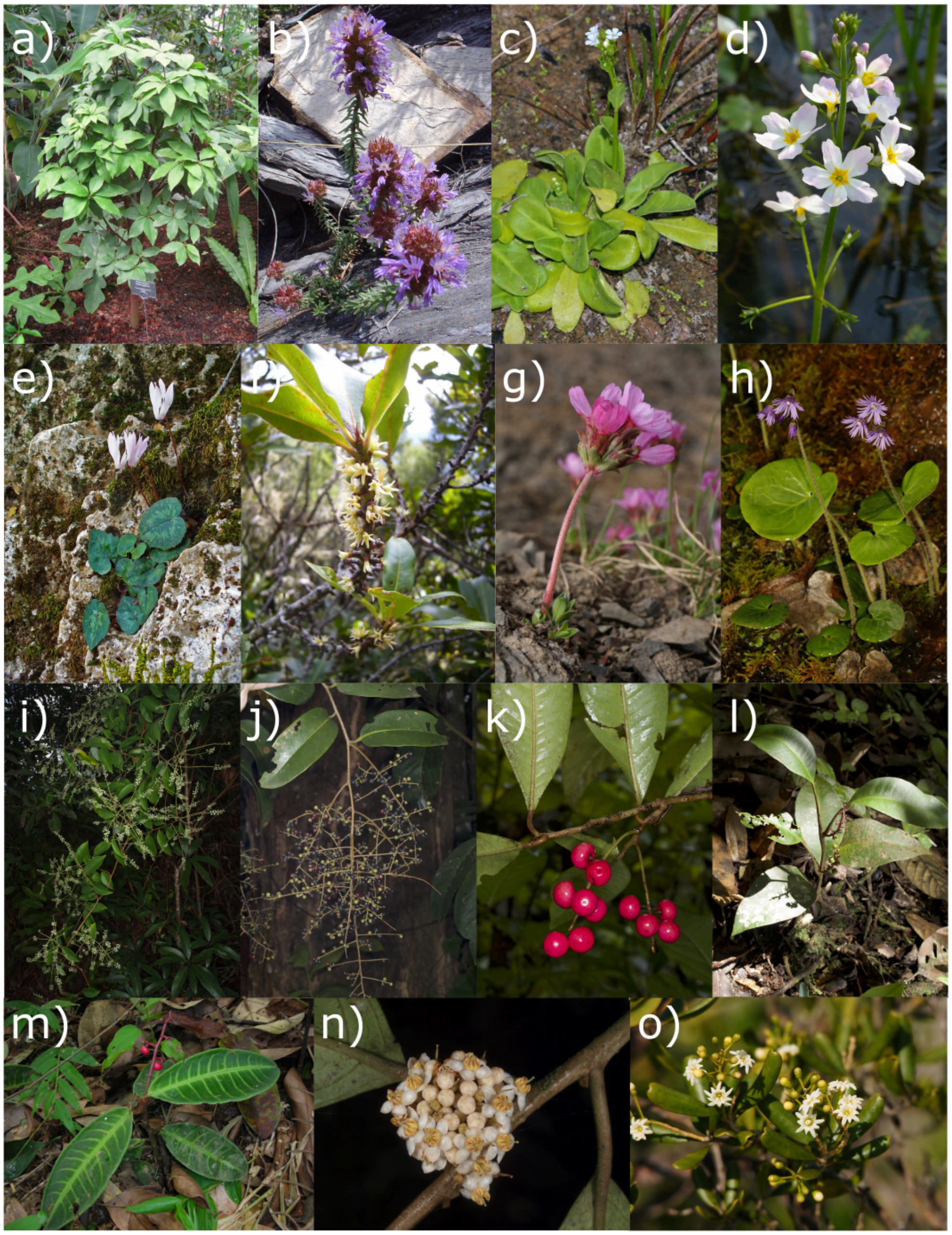


Figure 4-1. Examples of morphological diversity of Primulaceae, a) *Deherainia smaragdina*, b) *Coris monospeliensis*, c) *Samolus valerandi*, d) *Hottonia palustris* inflorescence, e) *Cyclamen creticum*, f) *Pleiomeris canariensis*, g) *Androsace laevigata*, h) *Soldanella villosa*, i) the liana habit of *Embelia ribes,* j) *Embelia imbricata* infructescence*,* k) *Ardisia sp*. subgenus *Tinus* lateral infructescence, l) the soboliferous shrublet habit and fruit of *Labisia longistyla*, m) the soboliferous shrublet habit and fruit of *Emblemantha urnulata,* n) axial inflorescence of *Ardisia fulinginosa*, a member of *Ardisia* subgenus *Pimlandra,* o) *Jacquinia keyensis* inflorescence. Photo attribution: a) Daderot public domain; b) Retama CC BY-SA 4.0; c-d) Christian Fischer CC BY-SA 3.0; e) H. Zell CC BY-SA 3.0; f) Winahwaru CC BY-SA 4.0; g) Walter Siegmund CC BY-SA 3.0; h) Cptcv CC BY-SA 2.5; i-n) T.M.A. Utteridge; o) Hans Hillewaert CC BY-SA 3.0.

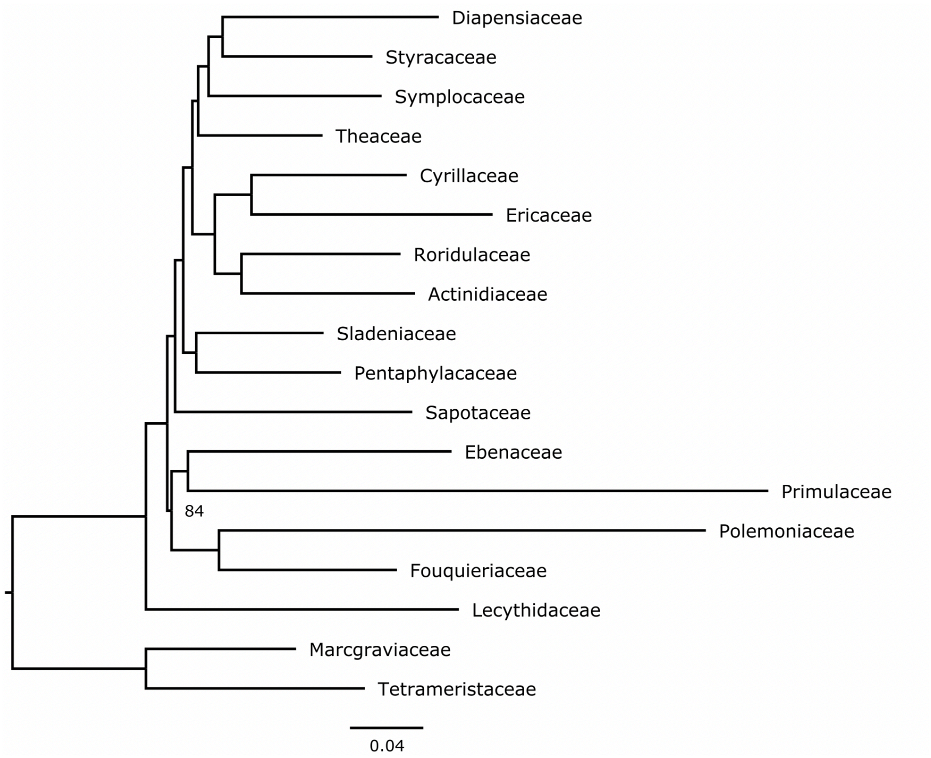
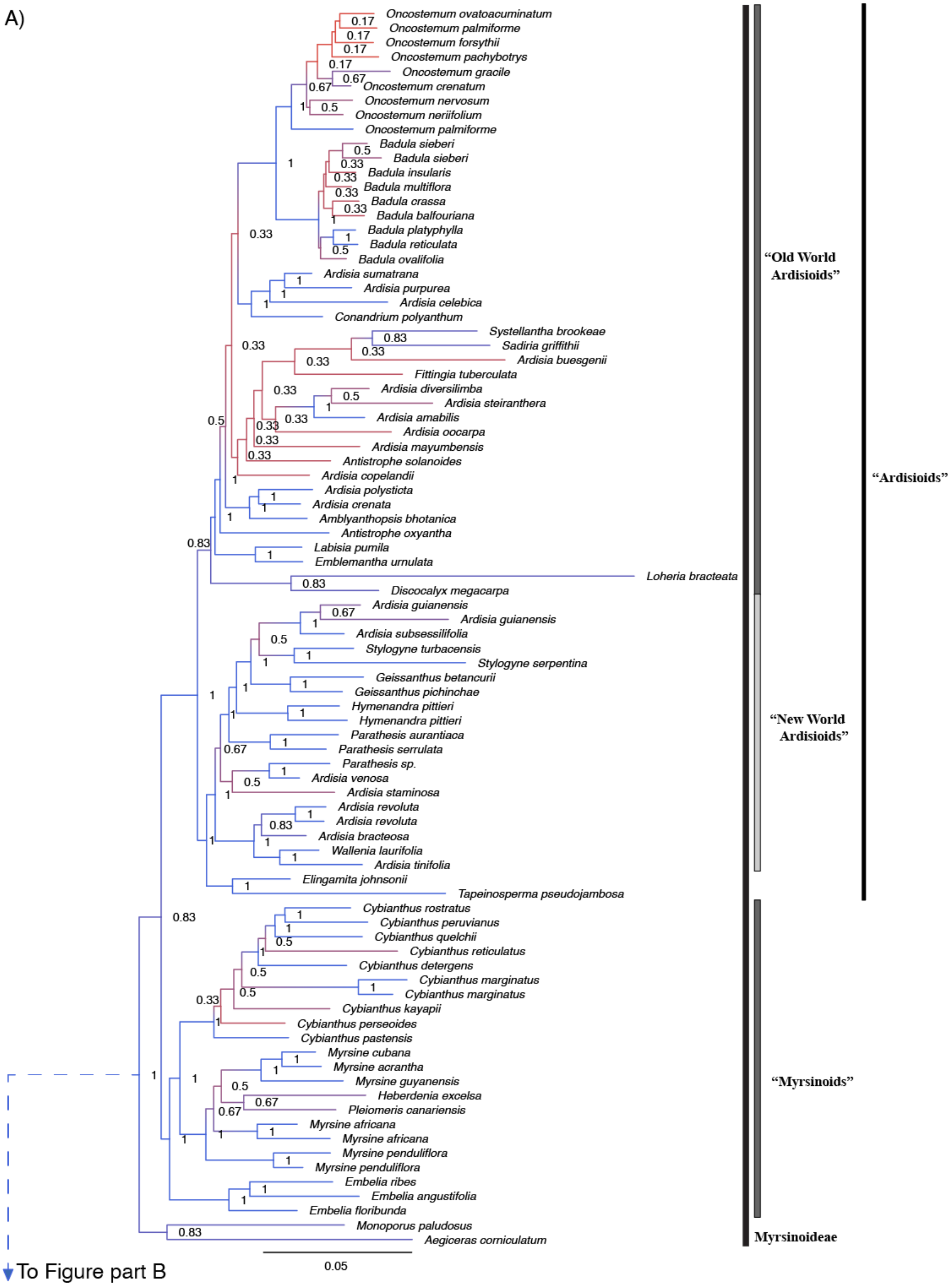


Figure 4-2. The phylogenetic relationships among families of Ericales recovered in the maximum likelihood tree produced with the Ericales1-exon dataset. All relationships had 100% ultrafast bootstrap support except the sister relationship between Primulaceae + Ebenaceae and Polemoniaceae + Fouquieriaceae, which received 84% support. The phylogeny was rooted with Marcgraviaceae + Tetrameristaceae, which with Balsaminaceae (not sampled), forms a clade that is sister to the rest of Ericales. By rooting in this way, the length of the branch separating this clade from the rest of the tree is not informative. Other branch lengths are estimated substitutions per site.


A)
Oncostemum ovatoacuminatum
0.17
Oncostemum palmiforme
0.17
Oncostemum forsythii
0.17
Oncostemum pachybotrys
0.17
Oncostemum gracile
0.67
Oncostemum crenatum
0.67
Oncostemum nervosum
1
0.5
Oncostemum neriifolium
Oncostemum palmiforme
Badula sieberi
0.5
Badula sieberi
1
0.33
Badula insularis
0.33
Badula multiflora
0.33
Badula crassa
0.33
Badula balfouriana
1
Badula platyphylla
1
Badula reticulata
0.5
0.33
Badula ovalifolia
Ardisia sumatrana
1
Ardisia purpurea
1
Ardisia celebica
1
Conandrium polyanthum
Systellantha brookeae
0.83
Sadiria griffithii
0.33
0.33
Ardisia buesgenii
0.33
Fittingia tuberculata
0.33
Ardisia diversilimba
0.5
Ardisia steiranthera
1
0.33
Ardisia amabilis
0.5
0.33
Ardisia oocarpa
0.33
Ardisia mayumbensis
0.33
Antistrophe solanoides
Ardisia copelandii
1
Ardisia polysticta
1
Ardisia crenata
1
1
Amblyanthopsis bhotanica
Antistrophe oxyantha
0.83
Labisia pumila
1
Emblemantha urnulata
Loheria bracteata
0.83
Discocalyx megacarpa
Ardisia guianensis
0.67
Ardisia guianensis
1
Ardisia subsessilifolia
0.5
Stylogyne turbacensis
1
1
Stylogyne serpentina
Geissanthus betancurii
1
1
Geissanthus pichinchae
1
Hymenandra pittieri
1
1
Hymenandra pittieri
Parathesis aurantiaca
1
0.67
Parathesis serrulata
Parathesis sp.
1
0.5
Ardisia venosa
1
Ardisia staminosa
Ardisia revoluta
1
0.83
Ardisia revoluta
1
Ardisia bracteosa
1
Wallenia laurifolia
1
Ardisia tinifolia
Elingamita johnsonii
1
Tapeinosperma pseudojambosa
0.83
Cybianthus rostratus
1
Cybianthus peruvianus
1
0.5
Cybianthus quelchii
1
Cybianthus reticulatus
0.5
Cybianthus detergens
Cybianthus marginatus
0.5
1
Cybianthus marginatus
0.33
Cybianthus kayapii
1
Cybianthus perseoides
Cybianthus pastensis
Myrsine cubana
1
Myrsine acrantha
1
1
1
Myrsine guyanensis
0.5
Heberdenia excelsa
0.67
0.67
Pleiomeris canariensis
Myrsine africana
1
1
1
Myrsine africana
Myrsine penduliflora
1
Myrsine penduliflora
Embelia ribes
1
Embelia angustifolia
1
Embelia floribunda
Monoporus paludosus
0.83
Aegiceras corniculatum
0.05
To Figure part B
“Old World Ardisioids”
“New World Ardisioids”
“Ardisioids”
“Myrsinoids”
Myrsinoideae

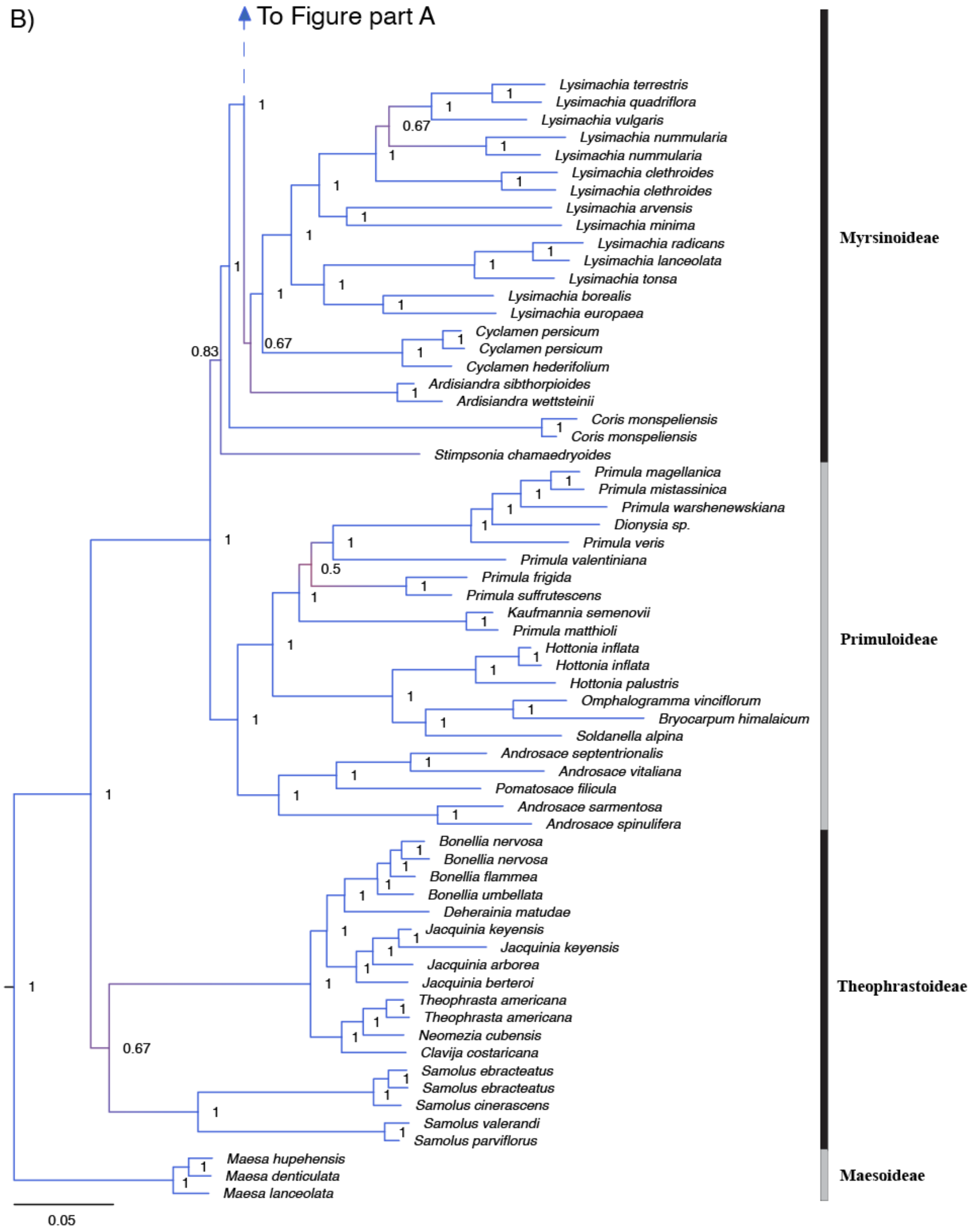


Figure 4-3. Best phylogenetic tree of the Primulaceae based on maximum likelihood analysis of the Ericales1-exon dataset (see text). A) The woody clade of Myrsinoideae, which comprises the Ardisioids, Myrsinoids, *Monoporus*, and *Aegiceras*. B) Relationships among the rest of Primulaceae. The topology for both is the Ericales1-exon-ML; values at nodes and branch colors correspond to the proportion of the six phylogenetic trees (the ML and ASTRAL trees for each of the three datasets) concordant with this topology. Branch lengths are in units of estimated substitutions per site. Outgroups not depicted.

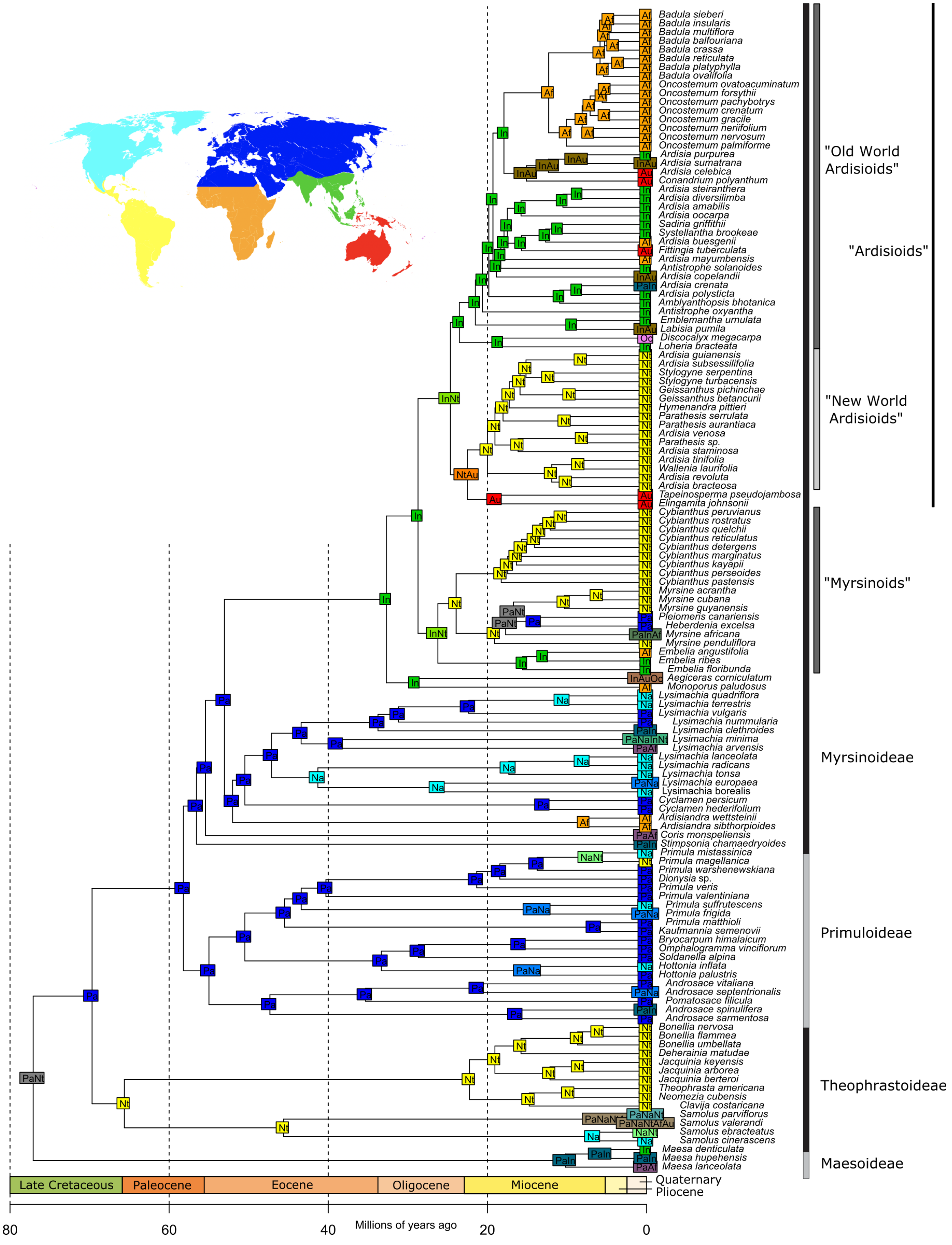


Figure 4-4. Results of the biogeographic analysis based on the topology from the Ericales1-exon-ML, reduced to a single tip per species. Node labels indicate the most likely state for that node in the DEC* model. Single area states (areas based on the biogeographic realms of Olson et al. 2001; see inset map): Nt = Neotropic (Yellow), Na = Nearctic (Light Blue), Pa = Palearctic (Dark Blue), Af = Afrotropic (Orange), In = Indo-Malay (Green), Oc = Oceania (Pink), Au = Australasia (Red). Map attribution to CarolSpears with modifications under the terms of the CC BY-SA 3.0 license.

Tables

Step 1	Remove all sequences flagged as potential paralogs by HybPiper or shorter than 25% of the average length of sequences for that gene
Step 2	Remove all sequences from poor-quality samples, including those that had fewer than 10 genes recovered at 25% of expected length based on the HybPiper target file
Step 3	Align with MAFFT
Step 4	Build gene trees with IQ-TREE
Step 5	Cut internal branches longer than 0.25 substitution per site, keeping only the largest subtree per gene. Cut terminal branches longer than three standard deviations above average for that gene tree
Step 6	Generate fastas based on the trimmed trees in Step 5
Step 7	Align fastas from Step 6 with MAFFT
Step 8	Proceed with using exon sequences for supermatrices, gene trees, and downstream analyses

Table 4-1. Summary of methods for generating and filtering the exon dataset for phylogenetic analyses.

Clade	Min age (Ma)	Max age (Ma)	Calibration type	Citation
Androsace	5.3	-	Fossil	Boucher et al. (2016)
Primula	16.0	-	Fossil	Boucher et al. (2016)
Lysimachia	28.1	-	Fossil	Boucher et al. (2016)
Crown Primulaceae	66.0	83.6	Fossil	Friis et al. (2010, 2021)
Stem Primulaceae	94.92	94.92	Secondary calibration	Magallón (2015)

Table 4-2. The estimated ages used to time-calibrate the Ericales1-exon phylogeny.

Node	Present study (95% HDI)	Rose et al. (2018) point estimate	Difference in point estimates
Crown Primulaceae	77.1 (76.3 - 77.7)	79.5	-2.4
Theophrastoideae	65.6 (65.0 - 66.3)	70.0	-4.4
Theophrasteae	22.3 (21.3 - 23.0)	20.6	1.7
Primuloideae	55.0 (54.1 - 55.8)	51.0	4.0
Myrsinoideae	56.6 (55.6 - 57.2)	53.1	3.5
Maesoideae	10.2 (9.2 - 10.8)	24.1	-13.9
Primuloideae + Myrsinoideae	58.2 (57.4 - 59.0)	57.6	0.6

Table 4-3. Divergence time estimations (in millions of years before present) for key nodes as compared to a recent dated phylogeny of Ericales (Rose et al., 2018; dates from Figure 3 of that study). Dates in paratheses are bounds of the 95% highest density interval (HDI) for dated bootstrap trees generated by resampling partitions and sites within partitions.

Herbaceous taxon (secondarily woody members in parentheses)	**Subfamily**
Primuloideae (some spp. secondarily woody subshrubs)	Primuloideae
Samolus (one sp. secondarily woody subshrub)	Theophrastoideae
Coris (only sp. secondarily woody subshrub)	Myrsinoideae
Lysimachia (some spp. secondarily woody shrubs, subshrubs)	Myrsinoideae
Cyclamen	Myrsinoideae
Stimpsonia	Myrsinoideae
Ardisiandra	Myrsinoideae

Table 4-4. Summary of the herbaceous lineages and those that appear to have evolved secondary woodiness *sensu* Lens et al. (2005) within the otherwise woody Primulaceae.

Chapter V

Conclusions and Future Directions

The preceding chapters have addressed a variety of phylogenetic and systematic questions at differing evolutionary scales. Evolutionary biology is a dynamic field and each study described here was thus conducted in slightly different, yet overlapping, historical contexts. In this chapter, I review some of the main conclusions drawn in this dissertation, describe some results in an updated context based on recently published work, and discuss directions for future work on the topics addressed in this dissertation.

Relationships among the families of Ericales

In Chapter II, my coauthors and I conducted a detailed investigation into the evolutionary relationships among the major clades of Ericales. While we show that relationships among the "core" families of the order can be resolved with confidence, we argue that there is not yet support for resolving a fully bifurcating phylogeny of the order (Chapter IV). The inferred relationships among several families differed depending on several factors including the criteria used to filter the genomic data, the alignment algorithm, and the tree-building method. We therefore chose to represent these contentious relationships as a hexatomy.

Polytomies are often distinguished as being either hard or soft. A soft polytomy is a non-bifurcating node leading to three or more branches on a phylogeny that is thought to be fully resolvable given more data. In contrast, a hard polytomy usually defined as one where the evolutionary history it represents comprises a series of divergence events so rapid or otherwise ambiguous that no series of bifurcations can accurately represent the history.

The type of polytomy reflected in the backbone of Ericales remains to be seen. Our analyses in Chapter II included a mixture of chloroplast and nuclear genes from transcriptomes. Many of our results placed Primulaceae sister to Ebenaceae, with Sapotaceae sister to those, especially when we used MAFFT to produce the alignments, rather than PRANK. This result agrees with most phylogenies of Ericales based on chloroplast genes. We address the backbone of Ericales again briefly in Chapter IV, this time using 353 nuclear genes, where we recovered an entirely different set of relationships that would ostensibly resolve the hexatomy.

I believe that a series of phylogenetic analyses focusing on the plastid genome of Ericales using full plastome data would be capable of resolving the history of ericalean plastome evolution with confidence. The plastid genome is thought to recombine in only very rare circumstances (Doyle, 2022; Walker et al., 2019). Therefore, we can have more confidence in representing this history as a single set of bifurcations. Though not addressed here specifically, the data generated during research for this dissertation is likely capable of resolving the Ericales plastid phylogeny. However, we know the situation is much more complicated for the nuclear genome. As we show in Chapter II, there is little doubt that nearly all Ericales have shared multiple rounds of whole genome duplication (WDG) in their history (Chen et al., 2020; Zhang et al., 2020, 2022). Extant species in the order hybridize with one another (Chapter III), and ancient lineages in the group may have as well (Stull et al., 2020).

Perhaps a fruitful line of investigation, to gain additional insight into relationships among the families of Ericales, will come from the comparison of sequenced genomes. As of February 2022, there are assembled genomes available for Theaceae, Ericaceae, Actinidiaceae, Sapotaceae, Sarraceniaceae, Balsaminaceae, Primulaceae, and Ebenaceae published on Genbank. Many of these have been published in the past two years, though not all are annotated in a way that makes them readily available for phylogenetic analyses. Comparing the order of genes (synteny) among annotated genomes may be a useful way to address specific questions that remain, such as whether Ebenaceae and Primulaceae are sister, and whether Sapotaceae is sister to those.

Whole genome duplications in Ericales

In Chapter II we also used phylotranscriptomics to investigate which clades of Ericales share whole genome duplications (WGDs). Our investigation was conducted contemporaneously with studies by other research groups focusing on slightly different questions, but that also shed light on ericalean WGDs. We found that the three nodes with the largest number of inferred gene duplications were those for 1) the non-Balsaminoid Ericales, 2) *Actinidia*, and 3) *Camellia*. Importantly, most methods used to investigate WGDs with transcriptomes have no way of distinguishing a duplication of *all* genes in the genome (i.e., a WGD) from a duplication of *many but not all* genes. For ancient WGDs, we expect that many duplicate genes will be lost and return to a single copy state during a process sometimes called “diploidization”.

An *Actinidia*-specific WGD, often called *Ad*-α, has been repeatedly confirmed by studies using whole genomes (Wu et al., 2019), and we interpreted the many gene duplications we observed in this genus as evidence for the *Ad*-α WGD. Wei et al. (2018) generated a genome of *Camellia* and reported two WGDs for that genus, one that was “recent” and one that was more ancient. Given the large number of gene duplications and peaks in K_s plots, we interpreted our results as showing evidence of a WGD specific to *Camellia*, which we named *Cm*-α for clarity.

We interpreted the many gene duplications along the backbone of Ericales as the older WGD reported in *Actinidia* and *Camellia*, and the same WGD that Huang et al. (2013) named *Ad*-β. An improved assembly of a *Camellia* genome after publication of Chapter II showed that the genus did not have a recent lineage-specific WGD, but rather contained many duplicated genes due to tandem gene duplications (Chen et al., 2020). These duplicated genes, and potentially problems accurately assembling them, may have caused a misleadingly large signal of many recently duplicated genes at that node in our results, which we interpreted as evidence of a *Camellia*-specific WGD. A recently published study by Zhang et al. (2022) also found that there is not a *Camellia*-specific whole genome duplication, but does support our finding that the *Ad*-β WGD is shared by most families of Ericales.

It is clear that there have been many WGDs that have shaped the evolution of angiosperms. It is also clear that there is room to improve our ability to detect WGDs, and especially our ability to differentiate what is and what is not a WGD. For Ericales, there now seems to be solid evidence for *Ad*-α (*Actinidia*-specific) and *Ad*-β (shared by all Ericales except perhaps the balsaminoid clade). The methods used by Leebens-Mack et al. (2019) inferred that there were two WGDs in sister clades along the backbone of Ericales, possibly due to the topology of their phylogeny. These inferred duplications now appear to both be *Ad*-β (Zhang et al., 2020). There is now also solid evidence that the node with the third most inferred gene duplications in Chapter II does not correspond to a WGD in Theaceae. Therefore, it seems entirely plausible, based on our results, that *Ad*-α and *Ad*-β are the only two Ericales-specific WGDs shared by entire ericalean families. However, Zhang et al. (2020) recently proposed lineage-specific WGDs in Styracaceae, Pentaphylaceae, Sapotaceae, Fouquieriaceae, and

Balsaminaceae. Whether researchers are overestimating the number of whole genome duplications that have occurred in angiosperms deserves careful consideration going forward.

Phylogeny of Lecythidaceae and admixture among species

In Chapter III we show that three species of *Eschweilera* hybridize with one another, though the extent to which this occurs and the effect that admixture has had on the ecology and evolution of the group remain to be seen. Given the results of our study and the extensive nuclear-plastome phylogenetic discordance we are observing in ongoing work on the family, I believe future studies will show that admixture is extensive among species of *Eschweilera*. It is now well-understood that *Lecythis* and *Eschweilera* are not monophyletic as currently circumscribed (Huang et al., 2015; Mori et al., 2017). My colleagues and I are currently preparing on a forthcoming re-circumscription of all genera of Neotropical Lecythidaceae so that all are monophyletic to the greatest extent possible (O. Vargas et al., *in prep*).

If admixture is widespread among the Neotropical Lecythidaceae, I believe this group could become a model system for studying the genomics of adaption in tropical trees. *Eschweilera coriacea* exhibits population structure across its extensive range (Chapter III), hybridizes with *E. wachenheimii* in the central Amazon, and is thought to hybridize with other species in French Guiana including *E. decolorans* and *E. sagotiana* (Schmitt et al., 2021). *Eschweilera* genomes may therefore have been shaped in complex ways by both locality-specific introgression and divergent selective pressures. It will also be important to attempt to quantify phenotypic and functional differences between populations of these species. Learning more about admixture in tropical trees should also lead to important conversations about what it means to conserve tropical biodiversity, since it is becoming increasingly clear using the species as a unit of conservation may ignore much of the genomic diversity that exists.

Phylogeny and systematics of Primulaceae

In Chapter IV, we describe the first phylogenetic analysis for nearly all Primulaceae using phylogenomic-scale nuclear data. As a result of this work, we now have a confident understanding of the major clades in the family and hypotheses of how they came to occupy their present-day distributions. We analyzed data from all but six genera of the family and the major clades within the family were resolved with strong support across methodologies. The subfamilies Myrsinoideae and Primuloideae are sister, with Theophrastoideae and Maesoideae successively sister to those. The position of *Samolus* remains somewhat uncertain but is most likely sister to the other Theophrastoideae or else is sister to Myrsinoideae + Primuloideae. Within Myrsinoideae, major clades corresponded to major biogeographic areas more than they corresponded to currently circumscribed genera. Revising the taxonomy of Myrsinoideae is necessary and will likely require collaboration among many research groups as well as extensive phylogenetic and morphological/anatomical investigation. Broader sampling of species from the Pacific Islands, including the genus *Tapeinosperma,* should provide additional insight into the biogeographic history of the New World members of *Ardisia* and allied genera. Broader sampling of *Ardisia* and *Hymenandra* will also provide insight into whether multiple trans-oceanic dispersal events have occurred in the clade, or whether the current distributions of extant members of these clades can be explained by a single dispersal event to the Neotropics.

Target capture sequencing for plant systematics

In Chapters III and IV, we used target capture sequencing, also sometimes called target enrichment, or targeted sequence capture. This method of generating data for phylogenetics has, in many ways, revolutionized the way the field approaches plant molecular systematics. It is now possible to sequence hundreds of times more loci, without more effort than conducting Sanger sequencing for a few genes. However, this technology is not without trade-offs.

For one, although target capture "probe" or "bait" oligonucleotides are designed to target low-copy-number genes, multiple paralogs are sometimes sequenced across multiple species or within a single sample. If multiple sequences are assembled in one sample, we can assume that we have recovered two paralogs and either exclude both from the analysis or used tree-based methods to determine which to keep. However, in the absence of a suitable reference genome, I believe it is still difficult to have much confidence that we are completely accounting for the presence of paralogs or mis-assembled sequences in target capture studies. A second consideration is that many target capture studies, including Chapters III and IV of this dissertation, use "loci" consisting of several concatenated exons. This is because target capture probes are designed to target exon regions that are relatively conserved across species, which makes the method possible. However, without a closely related reference genome for each sample, we don't know how far apart these exons are from one another, nor do we know how likely they are to share the same evolutionary history (i.e., no recombination between them), an issue that is widely problematic across many phylogenetic methods (Springer and Gatesy, 2016). This relatively new type of sequence data is also not yet available as assembled sequences in public databases like GenBank, creating a substantial barrier to its discovery and re-use by other research groups.

The Plant and Fungal Tree of Life Project (PAFTOL) has, as of February 2022, sequenced 7514 angiosperm genera (55% of angiosperm genera) and 412 angiosperm families (99% of angiosperm families) using target capture and "universal probes" for angiosperms in just the last six years (Baker et al., 2021; https://treeoflife.kew.org). This and the plethora of other studies utilizing this technology show that, despite potential shortcomings, target capture can be used to collect multi-locus nuclear data for any plant. I believe that target capture will continue to be a valuable tool for plant systematics, but that we can develop better tools for generating and filtering target loci assemblies. In particular, the increasing availability of assembled genomes should provide a means to verify suspected paralogy issues and further investigate the consequences of using concatenated, coding regions of DNA for phylogenetics.

Concluding thoughts

It is an exciting time to study the evolutionary history of plants. It seems possible that in the next few decades, long-read DNA sequencing may again revolutionize the way we generate phylogenetic datasets and change how we ask questions about evolutionary relationships. I believe that the cost of sequencing a draft genome will eventually become low enough that we will routinely sequence and assemble draft nuclear genomes for each of our samples, as we are now able to do for plastomes. This should allow us greater flexibility in our studies to choose the genomic regions we want to analyze. One could identify windows of conserved genomic regions to investigate deep relationships and other more quickly evolving regions of the same assemblies, or even the same genome alignment, for more recent divergences. This may also allow us to have more confidence in our ability to accurately align sequences and therefore reduce our reliance on exons for phylogenomics. Accurate, long sequence reads may help reduce the computational burden of assembling draft genomes, but assembling genomes, aligning assembled genomes, and accounting for inversions, large indels, and other structural changes will likely continue to be limited by computational capacity going forward (Hahn, 2019).

The more genomic data we attain, the clearer it becomes that the history of plant life is extremely complex. The research conducted for this dissertation has provided many new insights that improve our understanding of the evolutionary history of Ericales. And yet, like most scientific studies, we are left with more questions than definitive answers. As I believe this dissertation makes clear, there are innumerable future directions for studies of the single branch of life that is Ericales. It will be fascinating to see what we can learn about the forces that have led to the diversity of Ericales from the integration of genomic, phenomic, and other "-omic" insights, as well as machine learning, in the coming decades.

Systematists and other biologists must also use this information to amplify and prioritize efforts to conserve Earth's biodiversity in the face of climate change and other growing human impacts. The reasons for this are many—I will mention only a few here. For one, we simply do not know the potential of undescribed and underdescribed species to benefit human lives directly by improving our medicines, foods, and other material resources. We also cannot confidently predict the effects that widespread species loss will have on Earth's ecosystems, effects that are very likely to be irreversible in many cases. Finally, biological diversity has an incredible potential to enrich our lives in many ways if we take the time to appreciate it. The methods described in this dissertation offer a few lenses with which to view living things. There are also many other perspectives and ways of interacting with the non-human world that can inspire and amaze. Perhaps by reflecting on what we find most meaningful about the biological world, we can more holistically value its impact on our own lives and do more to act accordingly.

Appendix A

Supplementary Figures, and Tables for Chapter II

Table A-1. The origins of transcriptome assemblies, reference genomes, and raw reads used in homolog clustering. Citations associated with raw reads available on the National Center for Biotechnology Information Sequence Read Archive are included where such information was provided in the accession record or could be confidently identified through a Google Scholar query of the relevant accession information. Superscripts in taxon names are included to differentiate between samples from the same species and correspond to the same sample throughout.

Available at: https://dx.doi.org/10.7302/4153

Figure A-1. Phylogram inferred using maximum likelihood (ML) on a supermatrix consisting of the genes *rpoC2*, *rbcL*, *nhdF*, and *matK*. The topology of the tree was used to assign taxa into eight groups for hierarchical homolog clustering, such that each clustering group was monophyletic and not prohibitively large.

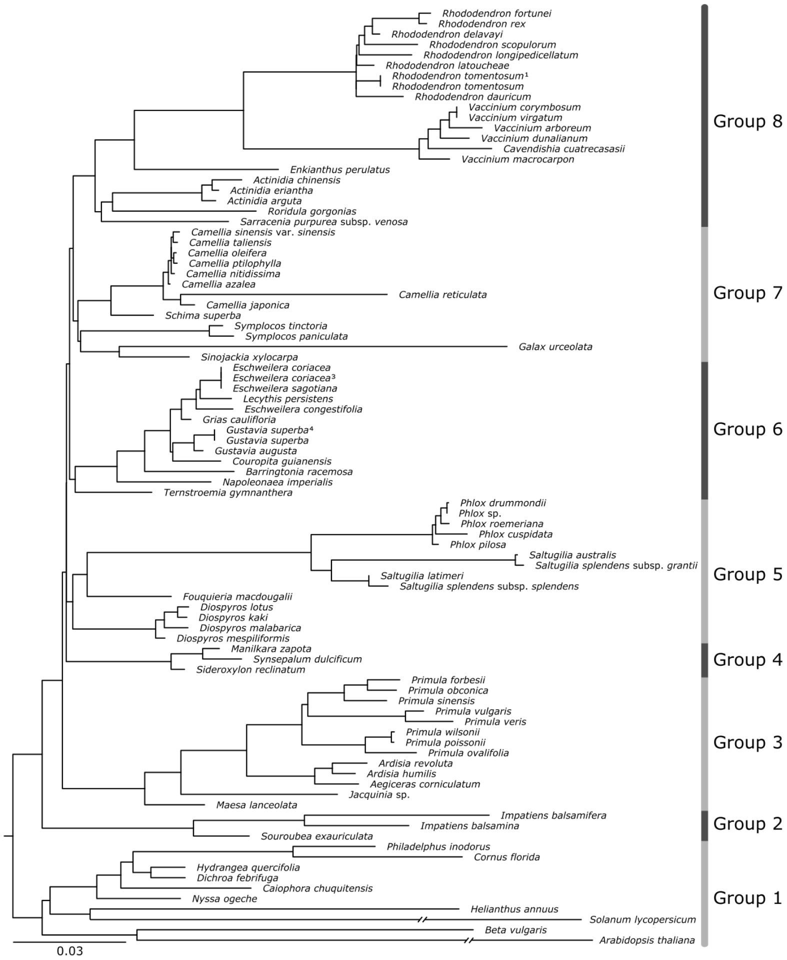


Figure A-2. Results of a two-topology test comparing gene- wise support in the 387 ortholog data set for two alternative placements of Lecythidaceae recovered by the maximum likelihood (ML) search. Delta gene-wise log-likelihood represents the extent to which an ortholog supports the ML topology over that recovered unanimously in rapid bootstraps. The dashed line represents the cumulative difference in log-likelihood between the two competing topologies, such that removing any of the 27 orthologs with a score more positive than that would likely cause the ML topology to change.

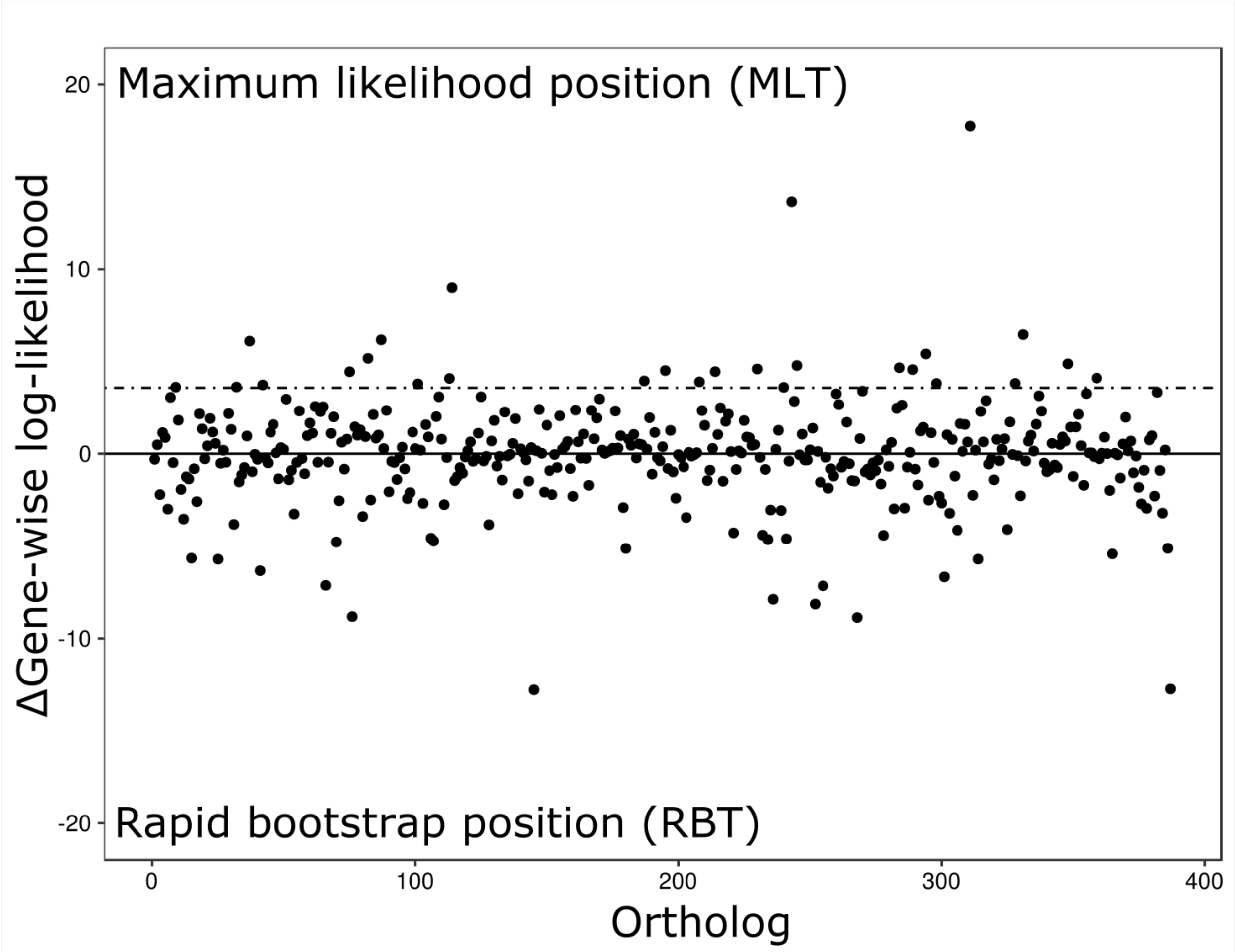


Table A-2. Results of unguided, regular bootstrapping with the 387 ortholog supermatrix in RAxML. The number of bootstrap replicates in which various contentious backbone relationships were recovered are reported.

Available at: https://dx.doi.org/10.7302/4153

Figure A-3. The maximum quartet support species tree topology for the 387 ortholog set generated with ASTRAL. Node support values are ASTRAL local posterior probabilities and nodes receiving support less than 1.0 are labeled. Branch lengths are in coalescent units.

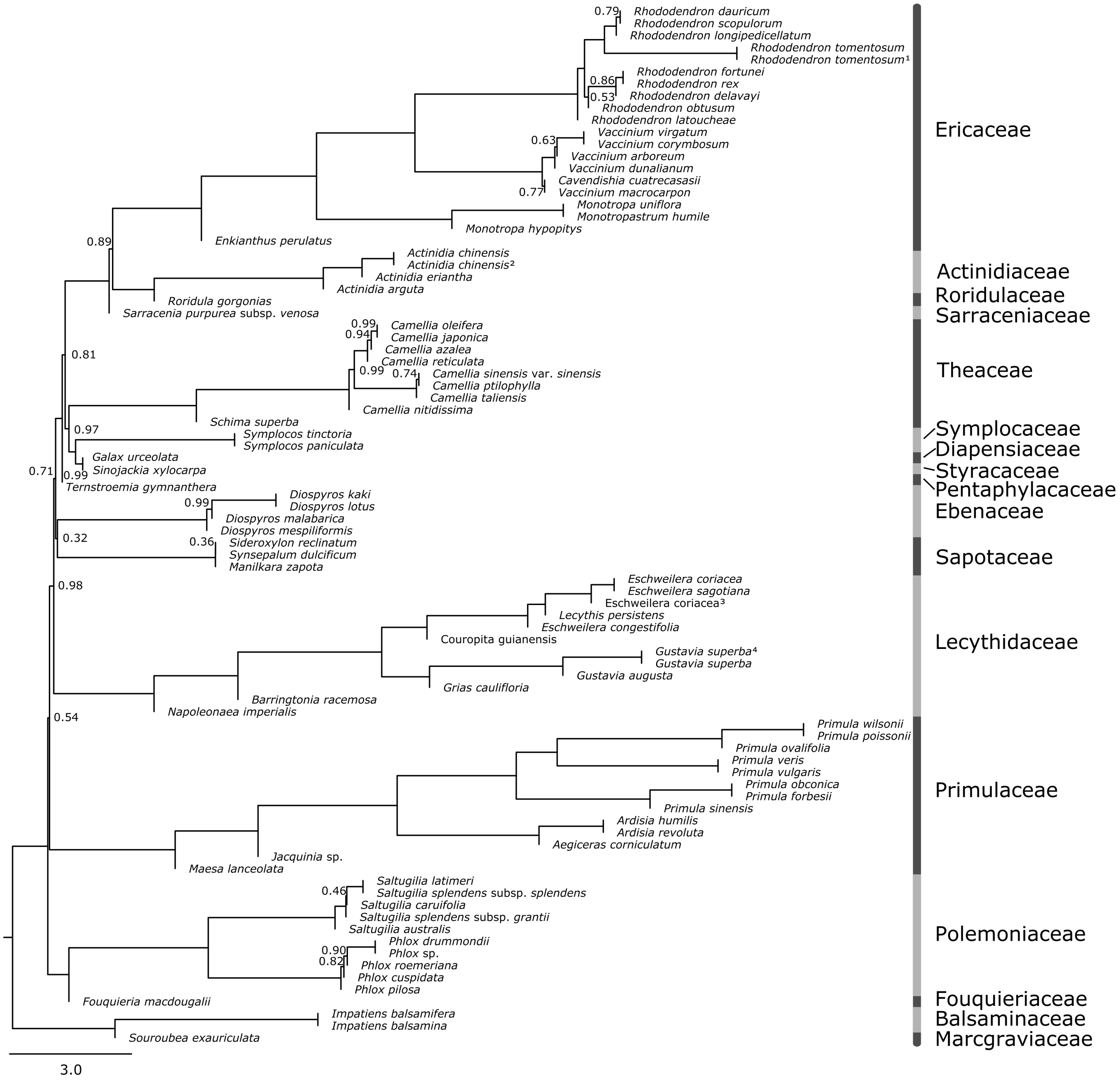


Figure A-4. Results from the gene-wise comparisons of likelihood contributions for the three candidate topologies for the 387 ortholog data set. (A) The absolute log-likelihood for each ortholog, sorted by which topology they best support and organized in descending order of likelihood. (B) The cumulative difference in log-likelihood relative to the ML topology as orthologs are added across the supermatrix. Delta gene-wise log-likelihood (ΔGWLL) represents the extent to which the topology has a worse score than the ML topology; values below zero indicate that a topology has a better likelihood than the ML topology. (C) Schematic of the three candidate topologies, the MLT (black), RBT (red), and AT (blue).

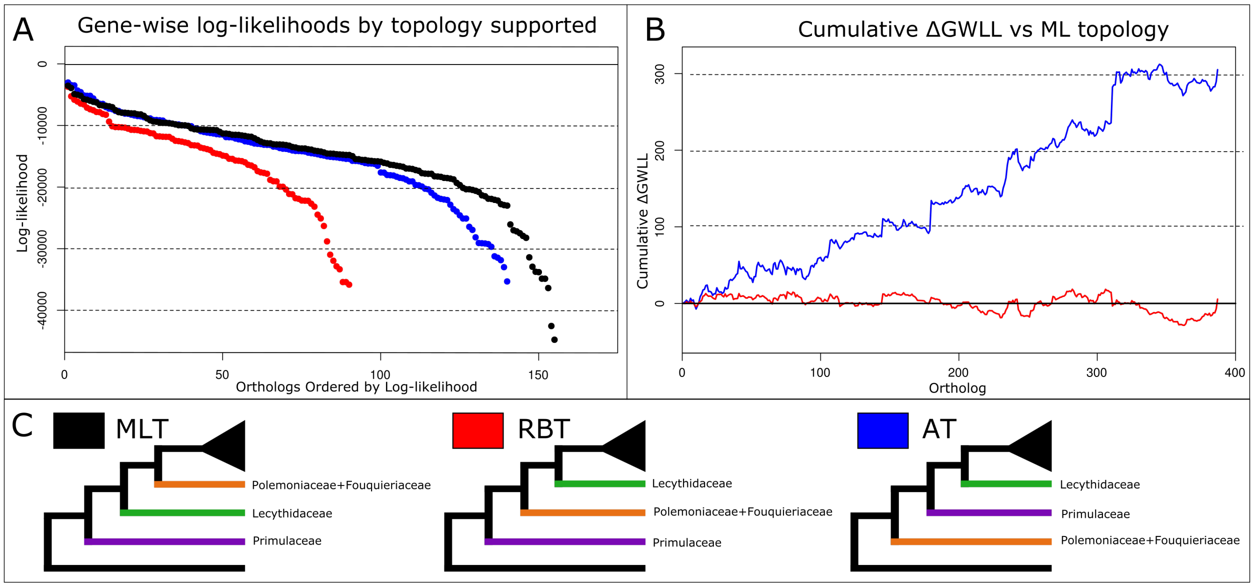


Table A-3. Log-likelihood penalties incurred by constraining contentious edges consistent with each of the three candidate topologies for a 387 ortholog data set using EdgeTest. Direct comparisons of scores between conflicting relationships can be used as a metric of support with lower scores suggesting stronger support.

Available at: https://dx.doi.org/10.7302/4153

Table A-4. Phyckle results regarding the placement of Ebenaceae for the 387 ortholog data set. Of the relationships examined, Ebenaceae + Sapotaceae was supported by the highest number of orthologs whether or not two log-likelihood support difference was required.

Available at: https://dx.doi.org/10.7302/4153

Figure A-5. Cladogram showing results from quartet sampling on the rapid bootstrap topology (RBT) from a 387 ortholog data set. Branch labels show quartet concordance (QC), quartet differential (QD), and quartet informativeness (QI), respectively, for each relationship. Quartet fidelity (QF) for each taxon is shown in parentheses after the relevant taxon label.

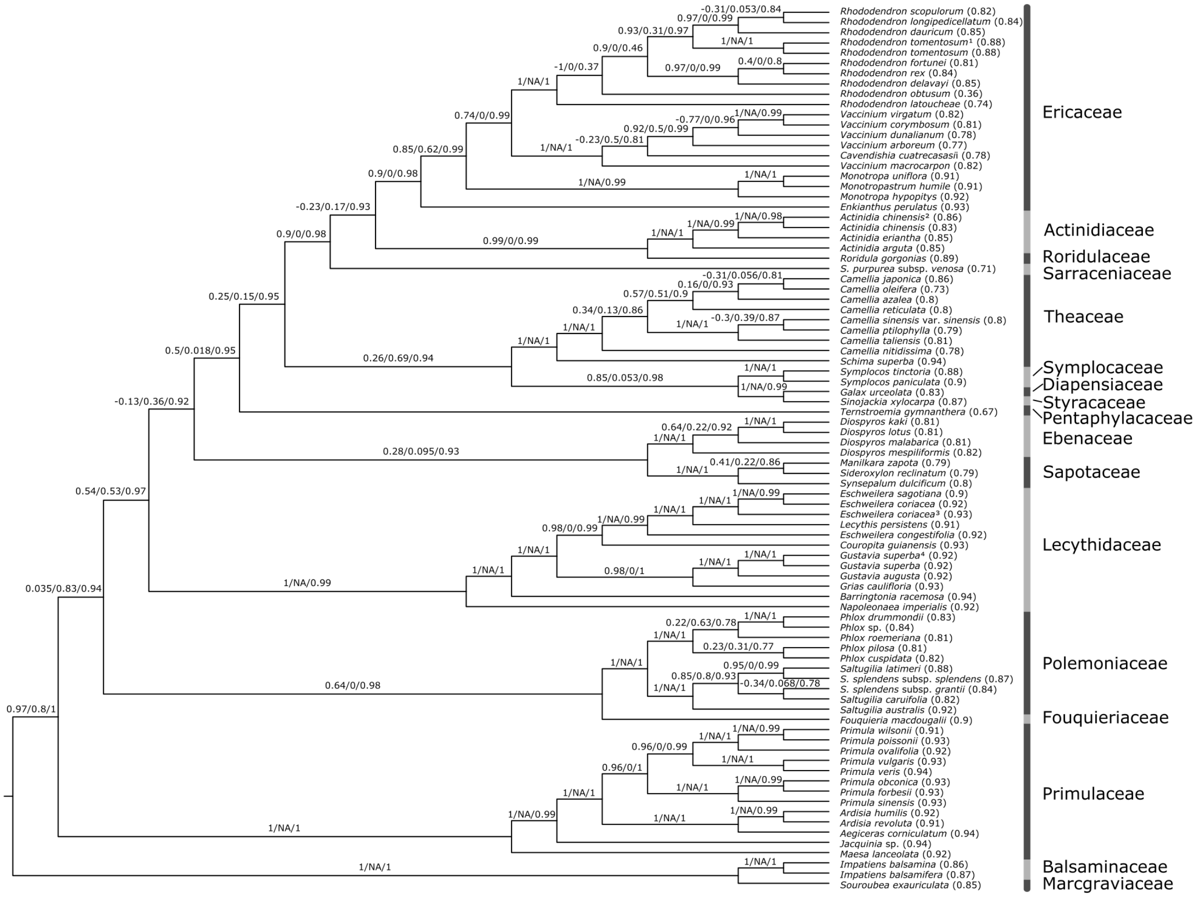


Figure A-6. (A) Ortholog tree concordance and conflict for the 387 ortholog set mapped to the rapid bootstrap topology. (B) The same analysis except requiring 70% ortholog tree bootstrap support for an ortholog to be considered informative. Blue indicates the proportion of informative orthologs that are concordant with the topology, green indicates the proportion of informative orthologs that support the single most common conflicting topology, red indicates all other informative ortholog conflict and gray indicates orthologs that are uninformative, either because of support requirements or lack of appropriate taxon sampling. Labels above and below branches indicate the number of informative orthologs that are concordant and in conflict with the branch respectively.

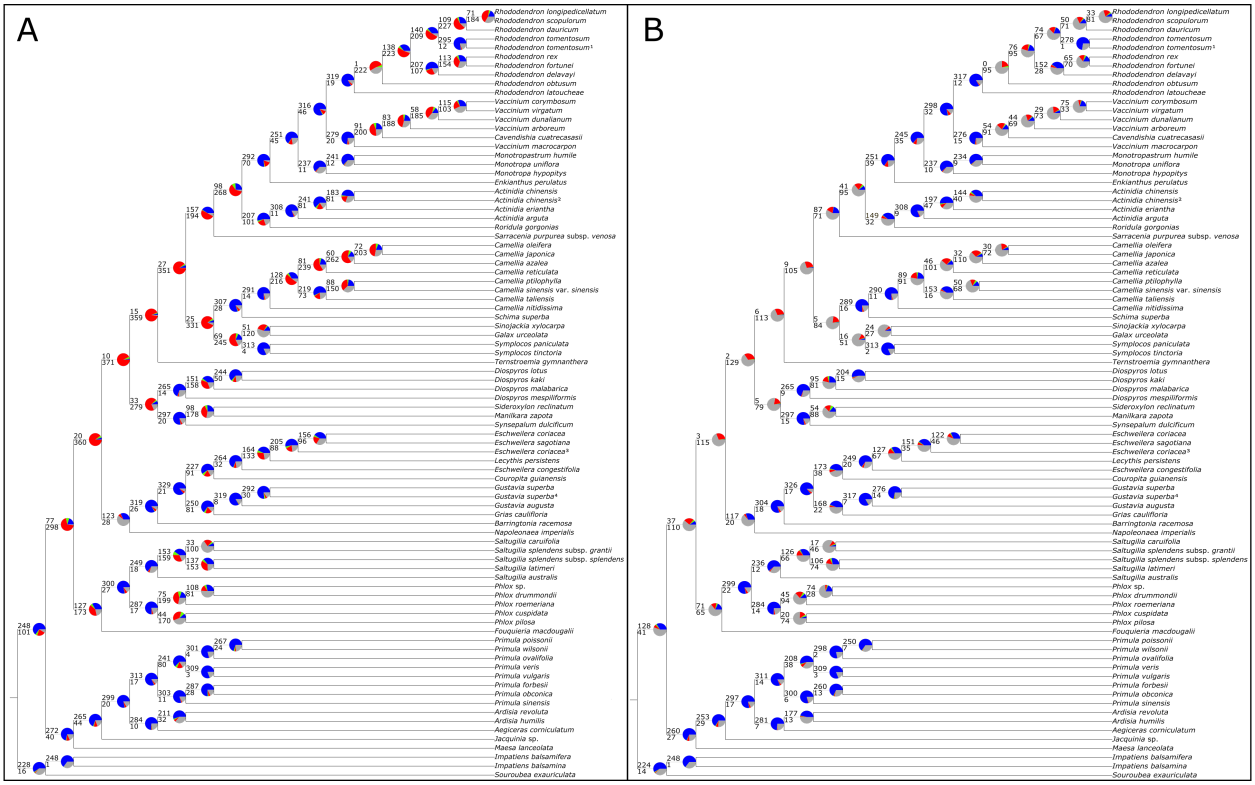


Table A-5. Phyckle results regarding the placement of Ebenaceae for the 2045 ortholog data set. Ebenaceae + Primulaceae was supported by the highest number of orthologs whether or not two log-likelihood support difference was required.

Available at: https://dx.doi.org/10.7302/4153

Figure A-7. Gene duplications mapped to a cladogram of the 449 ortholog MAFFT-aligned, partitioned, supermatrix ML tree—the single most commonly recovered species tree in this study. Number of inferred gene duplications is shown along branches. The diameter of circles at nodes are proportional to the number of duplications.

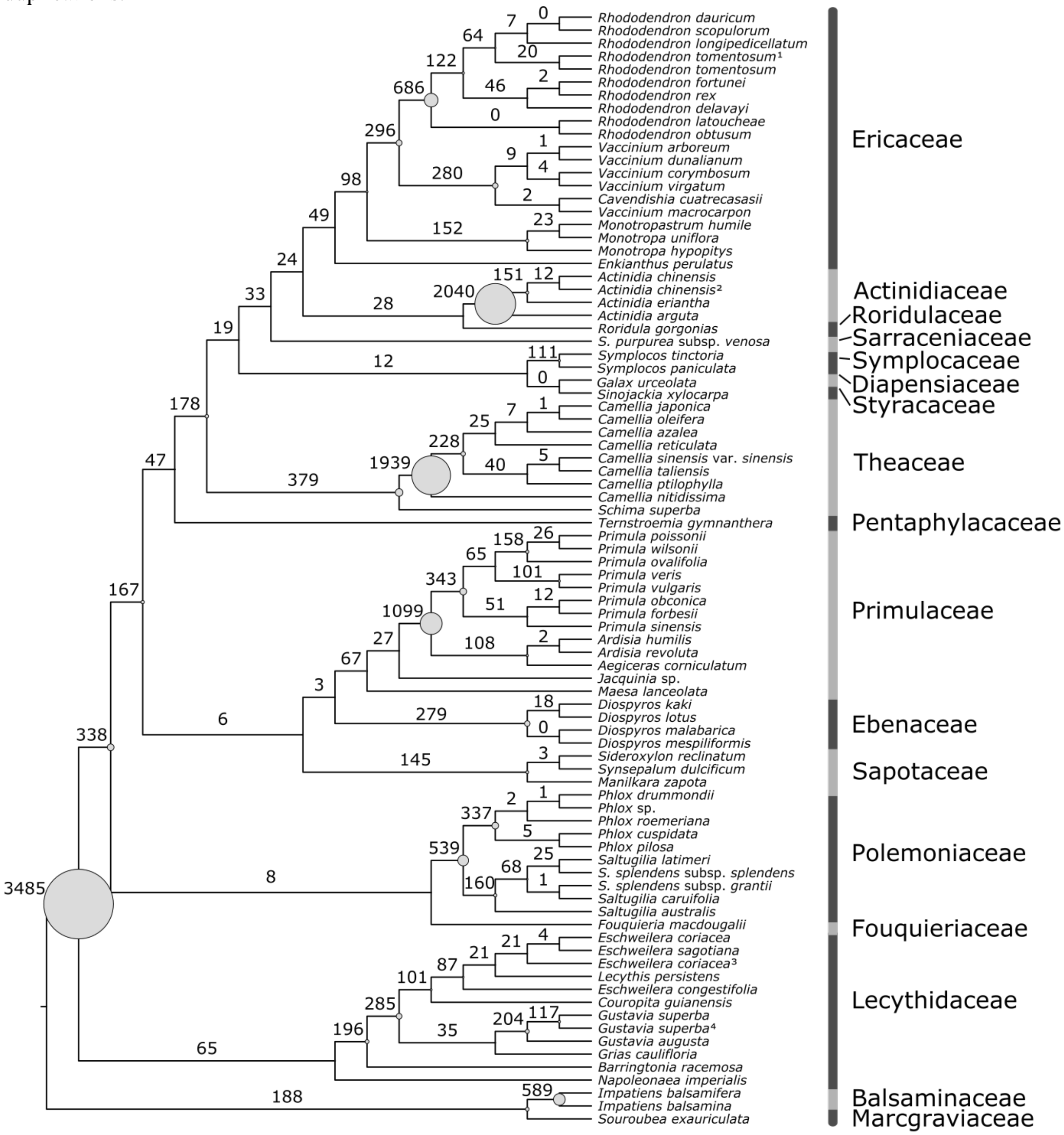


Figure A-8. Single species K_S plots for pairs of paralogs within each taxon as a density plot. Density peaks indicate evidence for a large proportion of genes having been duplicated at approximately the same time, as would occur during a whole genome duplication. For each taxon, the *x*-axis ranges from 0.0–3.0 with each tick representing 0.5 synonymous substitutions between paralogs. The *y*-axis is scaled to the maximum density value for each transcriptome and ticks correspond to 0.25, 0.5, 0.75, and 1.0 of the maximum density, respectively. Input transcriptomes have been filtered to remove short sequences and sample-specific duplications since these can represent transcript splice-site variants or errors in assembly.

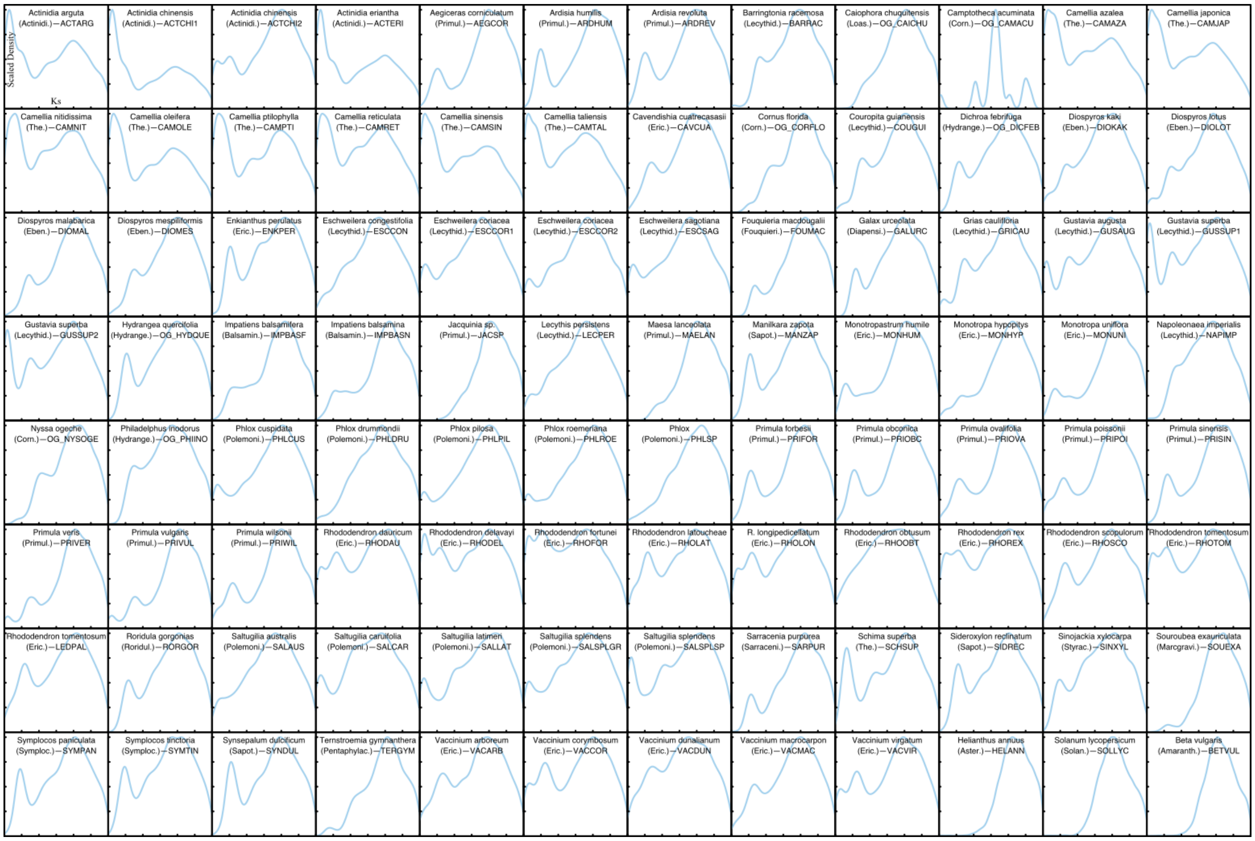


Figure A-9. Multispecies K_S plots representing a broad range of taxonomic pairings. Single species K_S density plots (i.e., pairs of paralogs) within each taxon are plotted on the same axis as a K_S density plot representing pairs of orthologs between the two taxa. The density peak for the orthologs corresponds to the time of divergence between the two taxa, because orthologs in both taxa begin accumulating synonymous substitutions after speciation. If the rate in the two taxa are equal and the accumulation of synonymous substitutions is clock-like, then the timing of the duplication events relative to divergence of the two taxa can be compared, with older events occurring farther to the right. The *x*-axis ranges from 0.0–3.0 with each tick representing 0.5 synonymous substitutions. The *y*-axis is scaled to the maximum density value of any of the three density distributions (therefore the magnitudes of all peaks are relative) and ticks correspond to 0.25, 0.5, 0.75, and 1.0 of the maximum density, respectively. Input transcriptomes have been filtered to remove short sequences and sample-specific duplications since because these can represent transcript splice-site variants or errors in assembly.

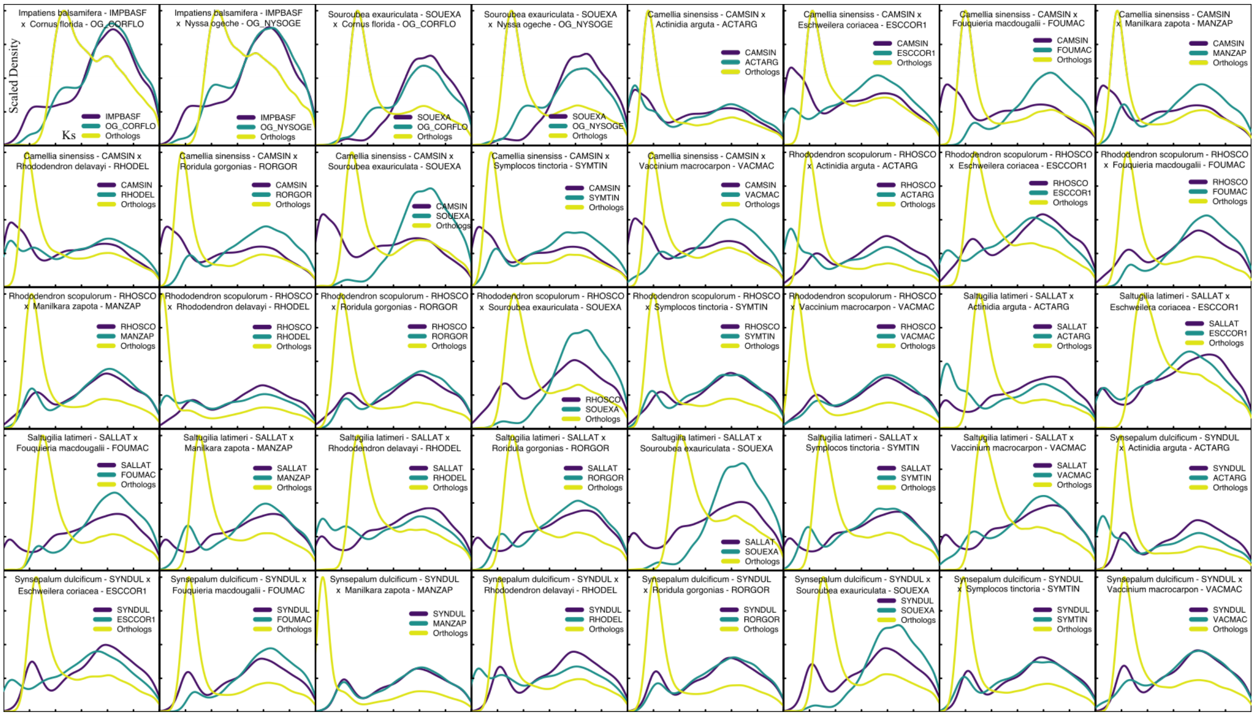
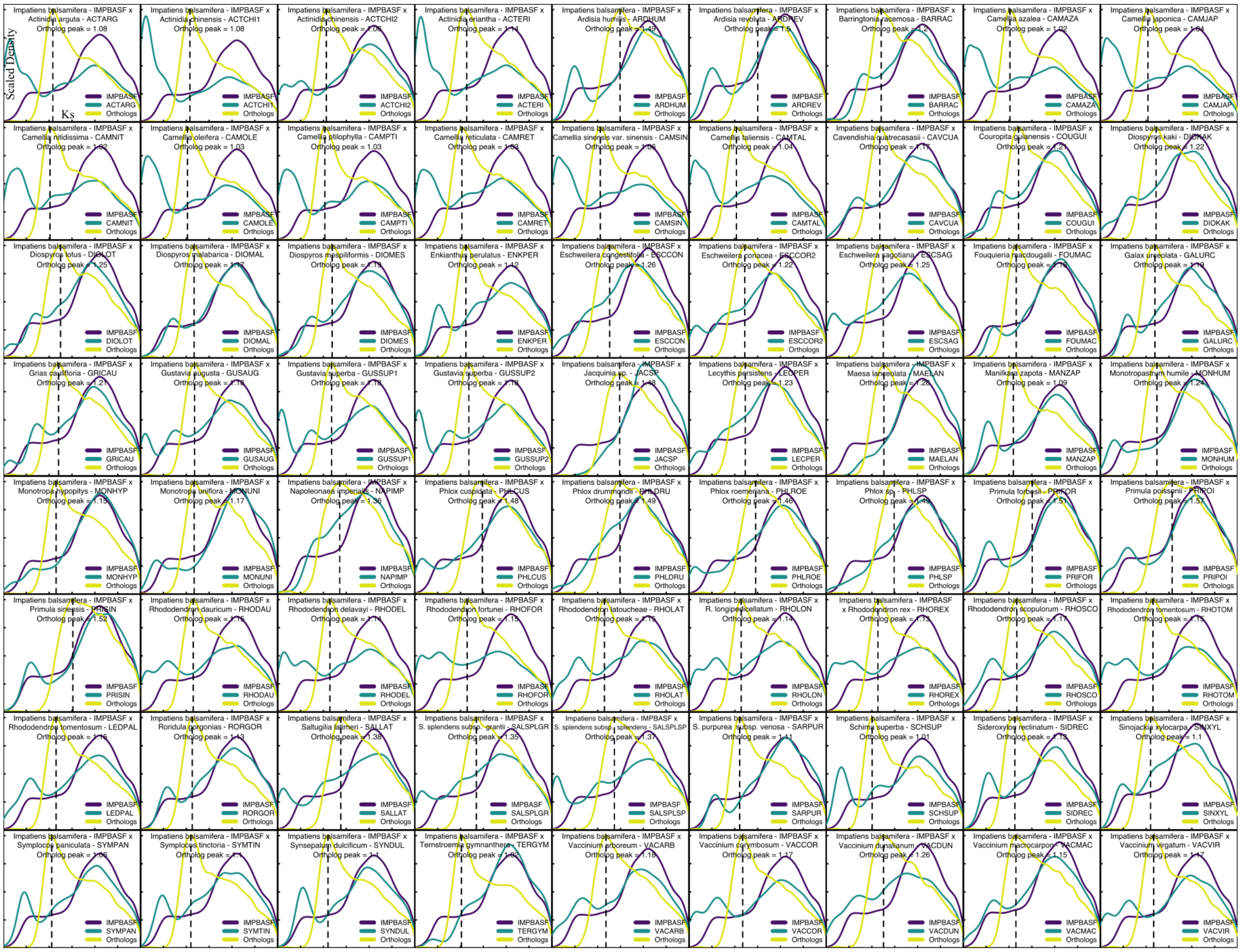


Figure A-10. Multispecies K_S plots representing each non-balsaminoid ericalean taxa paired with *Impatiens balsamifera* for which there is an unambiguous ortholog peak. Single species K_S density plots (i.e., pairs of paralogs) within each taxon are plotted on the same axis as a K_S density plot representing pairs of orthologs between the two taxa. The density peak for the orthologs corresponds to the time of divergence between the two taxa, because orthologs in both taxa begin accumulating synonymous substitutions after speciation. The dashed line represents the point along the *x*-axis where the ortholog peak achieves its maximum value and would be expected to occur at approximately the same point in all pairings under clock-like accumulation of synonymous substitutions since all pairs share the same MRCA. The *x*-axis ranges from 0.0–3.0 with each tick representing 0.5 synonymous substitutions. The *y*-axis is scaled to the maximum density value of any of the three density distributions (therefore the magnitudes of all peaks are relative) and ticks correspond to 0.25, 0.5, 0.75, and 1.0 of the maximum density, respectively. Input transcriptomes have been filtered to remove short sequences and sample-specific duplications because these can represent transcript splice-site variants or errors in assembly.



Appendix B

Supplementary Methods, Figures, Notes, and Tables for Chapter III

Figure B-1. Schematic of the process for making redeterminations based on all available evidence. The workflow was designed to allow specimen determinations to be updated to reflect new genetic evidence, as well as any remaining uncertainty, while minimizing any circularities that could result from using redeterminations made for some samples to update those of others. ***When considering whether a sample is sister to the rest, we only take into account other samples whose morphological determination is concordant with all available genetic evidence.

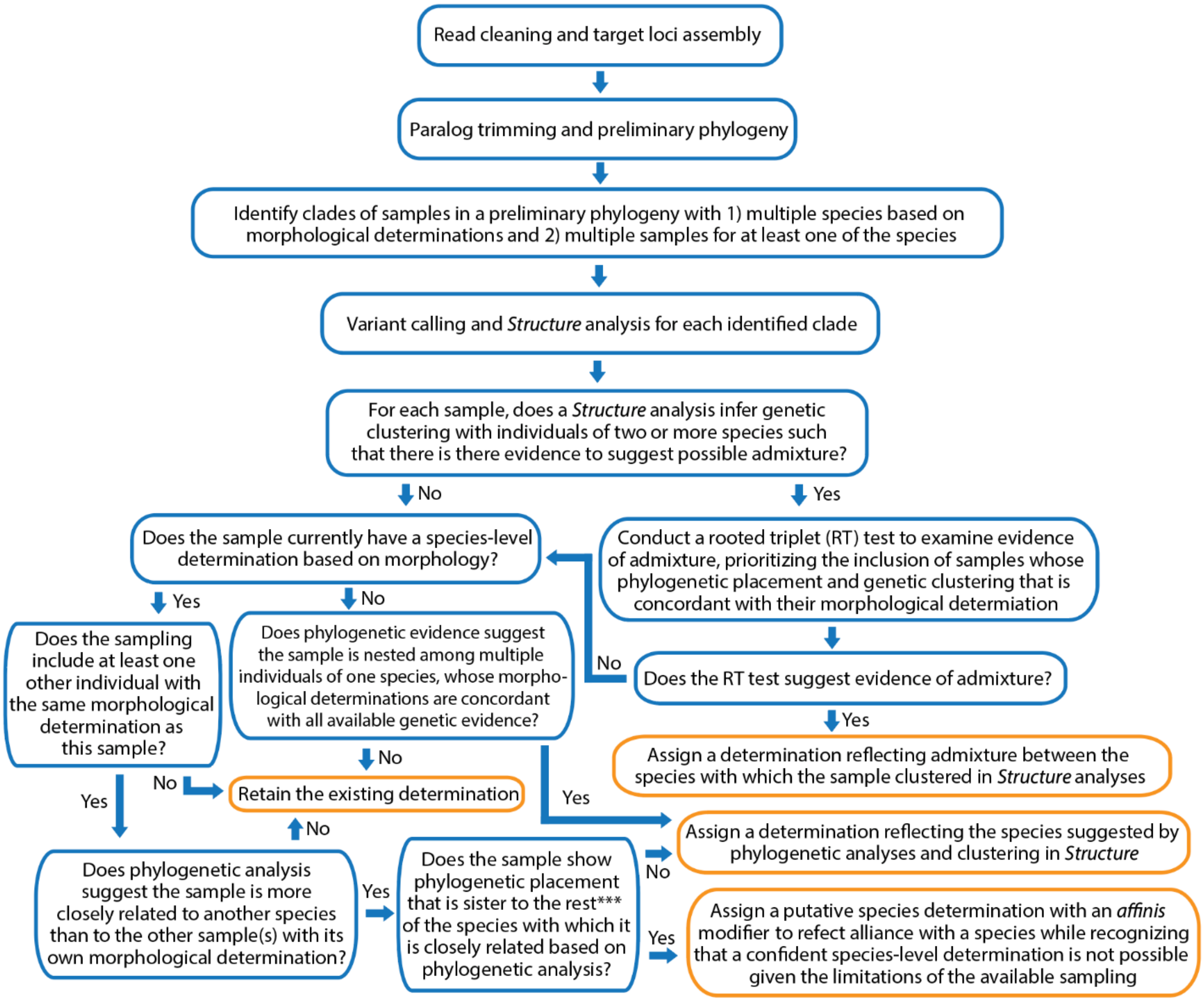

Read cleaning and target loci assembly
Paralog trimming and preliminary phylogeny
Identify clades of samples in a preliminary phylogeny with 1) multiple species based on morphological determinations and 2) multiple samples for at least one of the species
Variant calling and *Structure* analysis for each identified clade
For each sample, does a *Structure* analysis infer genetic clustering with individuals of two or more species such that there is there evidence to suggest possible admixture?
No
Yes
Does the sample currently have a species-level determination based on morphology?
Conduct a rooted triplet (RT) test to examine evidence of admixture, prioritizing the inclusion of samples whose phylogenetic placement and genetic clustering that is concordant with their morphological determiation
Yes
No
Does the sampling include at least one other individual with the same morphological determination as this sample?
Does phylogenetic evidence suggest the sample is nested among multiple individuals of one species, whose morpho-logical determinations are concordant with all available genetic evidence?
No
Does the RT test suggest evidence of admixture?
Yes
Assign a determination reflecting admixture between the species with which the sample clustered in *Structure* analyses
No
No
Retain the existing determination
Yes
Yes
Assign a determination reflecting the species suggested by phylogenetic analyses and clustering in *Structure*
No
Does phylogenetic analysis suggest the sample is more closely related to another species than to the other sample(s) with its own morphological determination?
Yes
Does the sample show phylogenetic placement that is sister to the rest*** of the species with which it is closely related based on phylogenetic analysis?
No
Yes
Assign a putative species determination with an *affinis* modifier to refect alliance with a species while recognizing that a confident species-level determination is not possible given the limitations of the available sampling

Figure B-2. The preliminary phylogeny of the Parvifolia clade estimated from a supermatrix of intron and exon target capture data. The phylogeny is shown as a cladogram and was generated with data from all 240 individuals, rooted on the genus *Napoleonaea* P.Beauv., then trimmed to include only individuals within the Parvifolia clade based on accepted taxonomy. Branch labels indicate RAxML rapid bootstrap support values. Tip labels are the accession codes used to represent each individual in all analysis files with species determinations in parentheses. In cases where a redetermination was made based on genetic evidence, the most recent morphological determination is noted on the left and the redetermination is on the right.

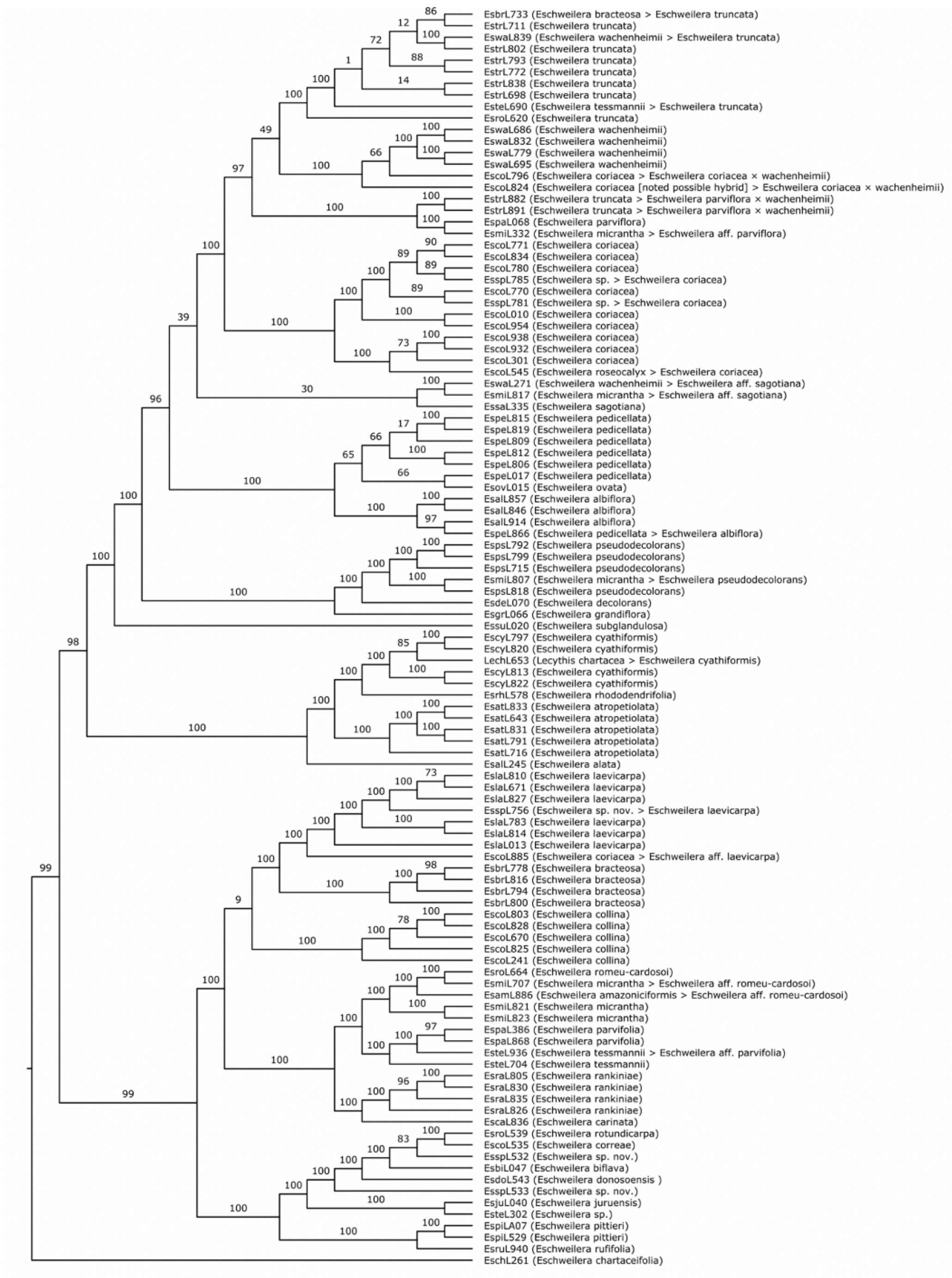

EsbrL733 (Eschweilera bracteosa > Eschweilera truncata)
EstrL711 (Eschweilera truncata)
EswaL839 (Eschweilera wachenheimii > Eschweilera truncata)
EstrL802 (Eschweilera truncata)
EstrL793 (Eschweilera truncata)
EstrL772 (Eschweilera truncata)
EstrL838 (Eschweilera truncata)
EstrL698 (Eschweilera truncata)
EsteL690 (Eschweilera tessmannii > Eschweilera truncata)
EsroL620 (Eschweilera truncata)
EswaL686 (Eschweilera wachenheimii)
EswaL832 (Eschweilera wachenheimii)
EswaL779 (Eschweilera wachenheimii)
EswaL695 (Eschweilera wachenheimii)
EscoL796 (Eschweilera coriacea > Eschweilera coriacea × wachenheimii)
EscoL824 (Eschweilera coriacea [noted possible hybrid] > Eschweilera coriacea × wachenheimii)
EstrL882 (Eschweilera truncata > Eschweilera parviflora × wachenheimii)
EstrL891 (Eschweilera truncata > Eschweilera parviflora × wachenheimii)
EspaL068 (Eschweilera parviflora)
EsmiL332 (Eschweilera micrantha > Eschweilera aff. parviflora)
EscoL771 (Eschweilera coriacea)
EscoL834 (Eschweilera coriacea)
EscoL780 (Eschweilera coriacea)
EsspL785 (Eschweilera sp. > Eschweilera coriacea)
EscoL770 (Eschweilera coriacea)
EsspL781 (Eschweilera sp. > Eschweilera coriacea)
EscoL010 (Eschweilera coriacea)
EscoL954 (Eschweilera coriacea)
EscoL938 (Eschweilera coriacea)
EscoL932 (Eschweilera coriacea)
EscoL301 (Eschweilera coriacea)
EscoL545 (Eschweilera roseocalyx > Eschweilera coriacea)
EswaL271 (Eschweilera wachenheimii > Eschweilera aff. sagotiana)
EsmiL817 (Eschweilera micrantha > Eschweilera aff. sagotiana)
EssaL335 (Eschweilera sagotiana)
EspeL815 (Eschweilera pedicellata)
EspeL819 (Eschweilera pedicellata)
EspeL809 (Eschweilera pedicellata)
EspeL812 (Eschweilera pedicellata)
EspeL806 (Eschweilera pedicellata)
EspeL017 (Eschweilera pedicellata)
EsovL015 (Eschweilera ovata)
EsalL857 (Eschweilera albiflora)
EsalL846 (Eschweilera albiflora)
EsalL914 (Eschweilera albiflora)
EspeL866 (Eschweilera pedicellata > Eschweilera albiflora)
EspsL792 (Eschweilera pseudodecolorans)
EspsL799 (Eschweilera pseudodecolorans)
EspsL715 (Eschweilera pseudodecolorans)
EsmiL807 (Eschweilera micrantha > Eschweilera pseudodecolorans)
EspsL818 (Eschweilera pseudodecolorans)
EsdeL070 (Eschweilera decolorans)
EsgrL066 (Eschweilera grandiflora)
EssuL020 (Eschweilera subglandulosa)
EscyL797 (Eschweilera cyathiformis)
EscyL820 (Eschweilera cyathiformis)
LechL653 (Lecythis chartacea > Eschweilera cyathiformis)
EscyL813 (Eschweilera cyathiformis)
EscyL822 (Eschweilera cyathiformis)
EsrhL578 (Eschweilera rhododendrifolia)
EsatL833 (Eschweilera atropetiolata)
EsatL643 (Eschweilera atropetiolata)
EsatL831 (Eschweilera atropetiolata)
EsatL791 (Eschweilera atropetiolata)
EsatL716 (Eschweilera atropetiolata)
EsalL245 (Eschweilera alata)
EslaL810 (Eschweilera laevicarpa)
EslaL671 (Eschweilera laevicarpa)
EslaL827 (Eschweilera laevicarpa)
EsspL756 (Eschweilera sp. nov. > Eschweilera laevicarpa)
EslaL783 (Eschweilera laevicarpa)
EslaL814 (Eschweilera laevicarpa)
EslaL013 (Eschweilera laevicarpa)
EscoL885 (Eschweilera coriacea > Eschweilera aff. laevicarpa)
EsbrL778 (Eschweilera bracteosa)
EsbrL816 (Eschweilera bracteosa)
EsbrL794 (Eschweilera bracteosa)
EsbrL800 (Eschweilera bracteosa)
EscoL803 (Eschweilera collina)
EscoL828 (Eschweilera collina)
EscoL670 (Eschweilera collina)
EscoL825 (Eschweilera collina)
EscoL241 (Eschweilera collina)
EsroL664 (Eschweilera romeu-cardosoi)
EsmiL707 (Eschweilera micrantha > Eschweilera aff. romeu-cardosoi)
EsamL886 (Eschweilera amazoniciformis > Eschweilera aff. romeu-cardosoi)
EsmiL821 (Eschweilera micrantha)
EsmiL823 (Eschweilera micrantha)
EspaL386 (Eschweilera parvifolia)
EspaL868 (Eschweilera parvifolia)
EsteL936 (Eschweilera tessmannii > Eschweilera aff. parvifolia)
EsteL704 (Eschweilera tessmannii)
EsraL805 (Eschweilera rankiniae)
EsraL830 (Eschweilera rankiniae)
EsraL835 (Eschweilera rankiniae)
EsraL826 (Eschweilera rankiniae)
EscaL836 (Eschweilera carinata)
EsroL539 (Eschweilera rotundicarpa)
EscoL535 (Eschweilera correae)
EsspL532 (Eschweilera sp. nov.)
EsbiL047 (Eschweilera biflava)
EsdoL543 (Eschweilera donosoensis)
EsspL533 (Eschweilera sp. nov.)
EsjuL040 (Eschweilera juruensis)
EsteL302 (Eschweilera sp.)
EspiLA07 (Eschweilera pittieri)
EspiL529 (Eschweilera pittieri)
EsruL940 (Eschweilera rufifolia)
EschL261 (Eschweilera chartaceifolia)
86
12
100
72
1
88
100
14
100
49
100
100
66
97
100
100
100
100
100
90
89
89
100
100
89
100
39
100
100
73
100
100
30
96
100
17
66
100
65
100
66
100
100
100
97
100
100
100
100
100
100
100
98
100
85
100
100
100
100
100
100
100
100
100
100
73
100
100
100
100
100
100
100
99
98
100
100
9
100
78
100
100
100
100
100
100
100
100
97
100
100
100
100
96
100
99
100
100
83
100
100
100
100
100
100
100
100

Figure B-3. Results of *Structure* analyses for alternative values of K when the run resulting in the best estimated probability of the data was not selected as the optimal K due to *a prior* taxonomic information. Each individual is labeled with a unique code used throughout all analyses and asterisks indicate samples from focal species collected at Reserve 1501. A) *Structure* results for the clade that included *E. romeu-cardosoi*, *E. carinata*, *E. micrantha, E. parvifolia*, E. *rankiniae*, and *E. tessmannii*. The estimated ln probability of the data was -5156.6, -5047.0, -5084.1, and -5171.0 for K=3, 4, 5, and 6 respectively. The optimal K was determined to be 6 because for lower values of K, the addition of more clusters tended to result in clustering that was increasingly concordant with the accepted taxonomy of Parvifolia clade based on previous morphological analyses. B) *Structure* results for the clade that included *E. atropetiolata*, *E. cyathiformis*, and *E. rhododendrifolia*. The estimated ln probability of the data was -3510.8.0, -3559.0, and -3483.7 for K=2, 3, and 4 respectively. The optimal K was determined to be K=3 because, though K=4 had a better scoring probability of the data, no individual was inferred to have more than 0.4% ancestry corresponding to the fourth cluster and K=3 was concordant with our *a priori* expectation given the current taxonomy of the Parvifolia clade.

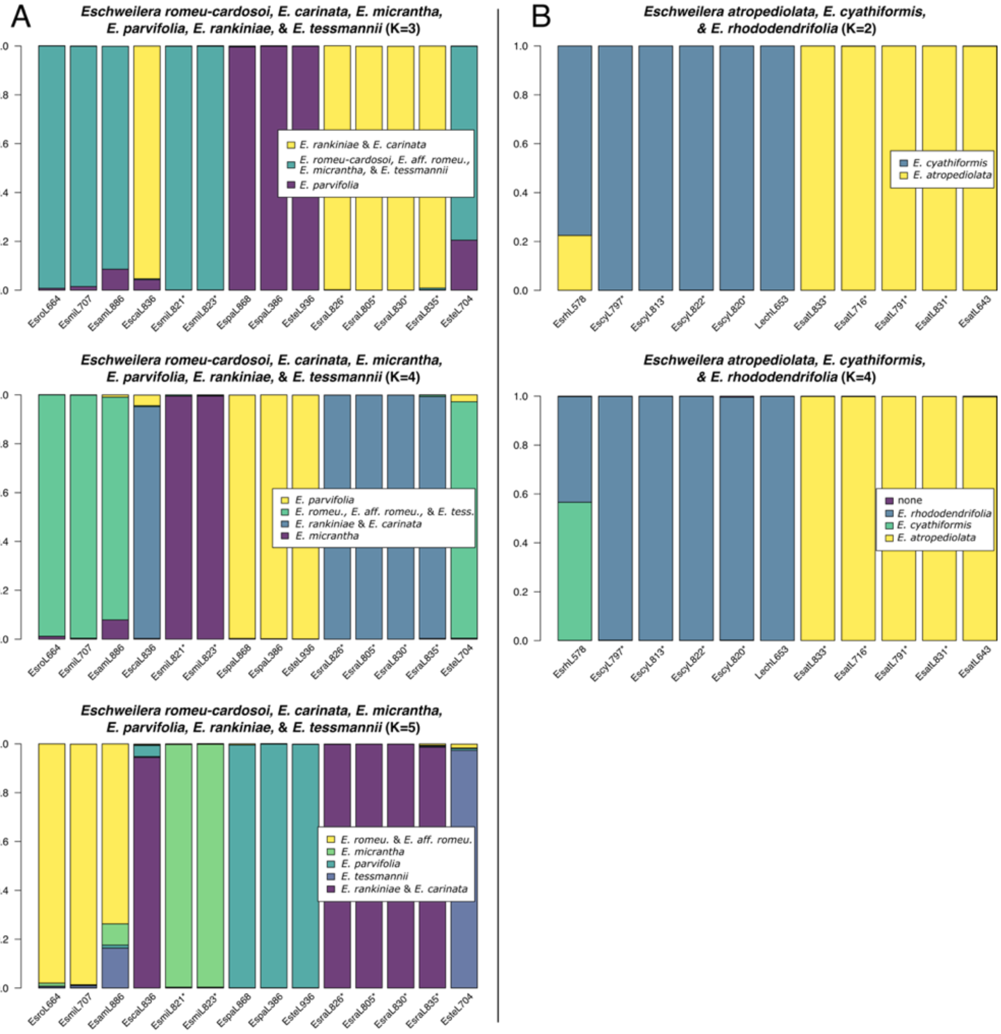


Figure B-4. Results of *Structure* analyses using a SNP dataset for the clade that included *E. coriacea*, *E. wachenheimii*, *E. sagotiana*, *E. truncata*, and *E. parviflora*. Each individual is labeled with a unique code used throughout all analyses and asterisks indicate samples from focal species collected at Reserve 1501. The estimated probability of the data was -11138.1 and -11061.8 for K=6 (A) and 7 (B) respectively. While the run with K=7 had the better score, the run with K=6 resulted in three individuals having nearly complete inferred ancestry from a single cluster that corresponded to *E. sagotiana*.

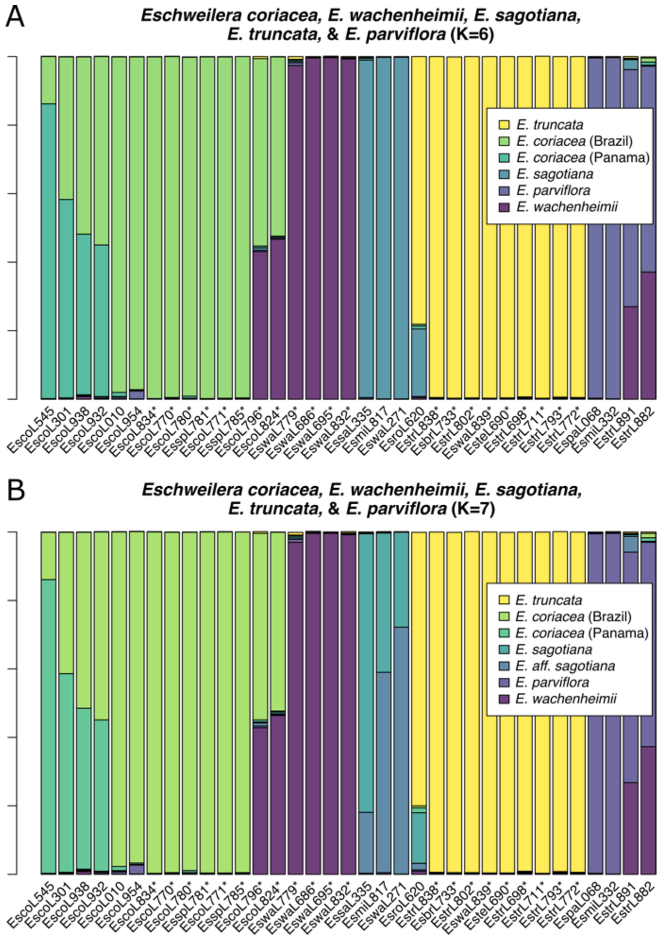


Figure B-5. Evidence of geographic structure among individuals of *E. coriacea*. The three-dimensional scatterplot shows the results of a genetic principal component analysis (PCA) that included all individuals of *E. coriacea* with no evidence of admixture. The x-, y-, and z-axes correspond to the first three principal components of the PCA, respectively. Each point represents an individual and colors correspond to the country in which it was collected.

Available at: https://dx.doi.org/10.7302/4153

Figure B-6. The Parvifolia phylogeny without reduced representation, produced using a supermatrix of intron and exon target capture data. Branch labels and coloration indicate concordance (1, blue) and conflict (0, red) with the results using an exon-only supermatrix with the same taxa and data filtering strategy. The tip labels represent taxon identities after redeterminations based on all available evidence. The phylogeny was rooted on an outgroup consisting of five members of the Integrifolia clade of *Eschweilera*.

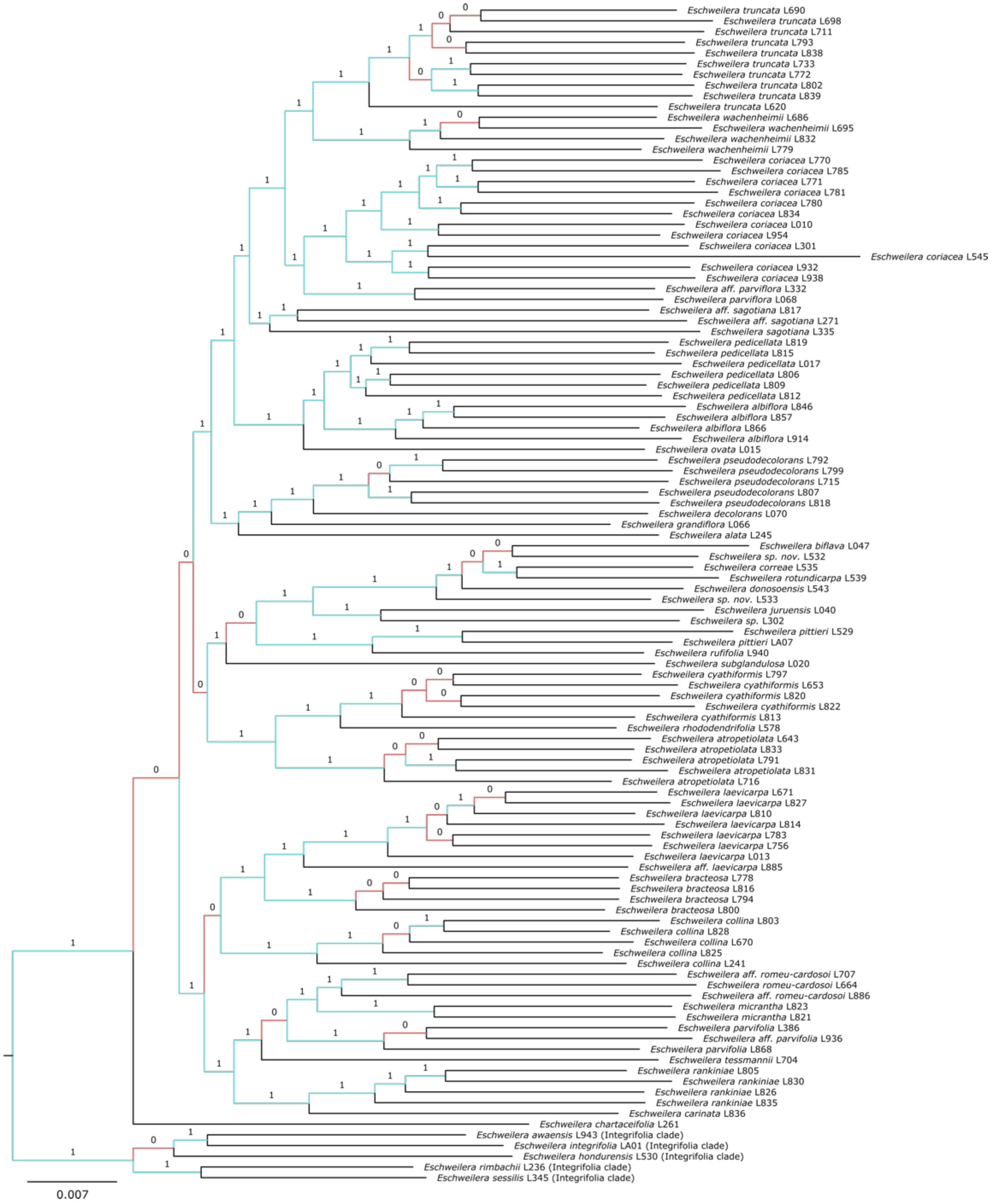

Eschweilera truncata L690
Eschweilera truncata L698
Eschweilera truncata L711
Eschweilera truncata L793
Eschweilera truncata L838
Eschweilera truncata L733
Eschweilera truncata L772
Eschweilera truncata L802
Eschweilera truncata L839
Eschweilera truncata L620
Eschweilera wachenheimii L686
Eschweilera wachenheimii L695
Eschweilera wachenheimii L832
Eschweilera wachenheimii L779
Eschweilera coriacea L770
Eschweilera coriacea L785
Eschweilera coriacea L771
Eschweilera coriacea L781
Eschweilera coriacea L780
Eschweilera coriacea L834
Eschweilera coriacea L010
Eschweilera coriacea L954
Eschweilera coriacea L301
Eschweilera coriacea L545
Eschweilera coriacea L932
Eschweilera coriacea L938
Eschweilera aff. parviflora L332
Eschweilera parviflora L068
Eschweilera aff. sagotiana L817
Eschweilera aff. sagotiana L271
Eschweilera sagotiana L335
Eschweilera pedicellata L819
Eschweilera pedicellata L815
Eschweilera pedicellata L017
Eschweilera pedicellata L806
Eschweilera pedicellata L809
Eschweilera pedicellata L812
Eschweilera albiflora L846
Eschweilera albiflora L857
Eschweilera albiflora L866
Eschweilera albiflora L914
Eschweilera ovata L015
Eschweilera pseudodecolorans L792
Eschweilera pseudodecolorans L799
Eschweilera pseudodecolorans L715
Eschweilera pseudodecolorans L807
Eschweilera pseudodecolorans L818
Eschweilera decolorans L070
Eschweilera grandiflora L066
Eschweilera alata L245
Eschweilera biflava L047
Eschweilera sp. nov. L532
Eschweilera correae L535
Eschweilera rotundicarpa L539
Eschweilera donosoensis L543
Eschweilera sp. nov. L533
Eschweilera juruensis L040
Eschweilera sp. L302
Eschweilera pittieri L529
Eschweilera pittieri LA07
Eschweilera rufifolia L940
Eschweilera subglandulosa L020
Eschweilera cyathiformis L797
Eschweilera cyathiformis L653
Eschweilera cyathiformis L820
Eschweilera cyathiformis L822
Eschweilera cyathiformis L813
Eschweilera rhododendrifolia L578
Eschweilera atropetiolata L643
Eschweilera atropetiolata L833
Eschweilera atropetiolata L791
Eschweilera atropetiolata L831
Eschweilera atropetiolata L716
Eschweilera laevicarpa L671
Eschweilera laevicarpa L827
Eschweilera laevicarpa L810
Eschweilera laevicarpa L814
Eschweilera laevicarpa L783
Eschweilera laevicarpa L756
Eschweilera laevicarpa L013
Eschweilera aff. laevicarpa L885
Eschweilera bracteosa L778
Eschweilera bracteosa L816
Eschweilera bracteosa L794
Eschweilera bracteosa L800
Eschweilera collina L803
Eschweilera collina L828
Eschweilera collina L670
Eschweilera collina L825
Eschweilera collina L241
Eschweilera aff. romeu-cardosoi L707
Eschweilera romeu-cardosoi L664
Eschweilera aff. romeu-cardosoi L886
Eschweilera micrantha L823
Eschweilera micrantha L821
Eschweilera parvifolia L386
Eschweilera aff. parvifolia L936
Eschweilera parvifolia L868
Eschweilera tessmannii L704
Eschweilera rankiniae L805
Eschweilera rankiniae L830
Eschweilera rankiniae L826
Eschweilera rankiniae L835
Eschweilera carinata L836
Eschweilera chartaceifolia L261
Eschweilera awaensis L943 (Integrifolia clade)
Eschweilera integrifolia LA01 (Integrifolia clade)
Eschweilera hondurensis L530 (Integrifolia clade)
Eschweilera rimbachii L236 (Integrifolia clade)
Eschweilera sessilis L345 (Integrifolia clade)
0.007

Figure B-7. Comparison of the reduced-representation Parvifolia phylogenies recovered with the two supermatrices. The Parvifolia phylogeny constructed using intron and exon target capture data is on the left and the exon-only Parvifolia phylogeny is on the right. Both phylogenies are presented as cladograms and the blue lines connecting tip labels indicate the same species in either tree.

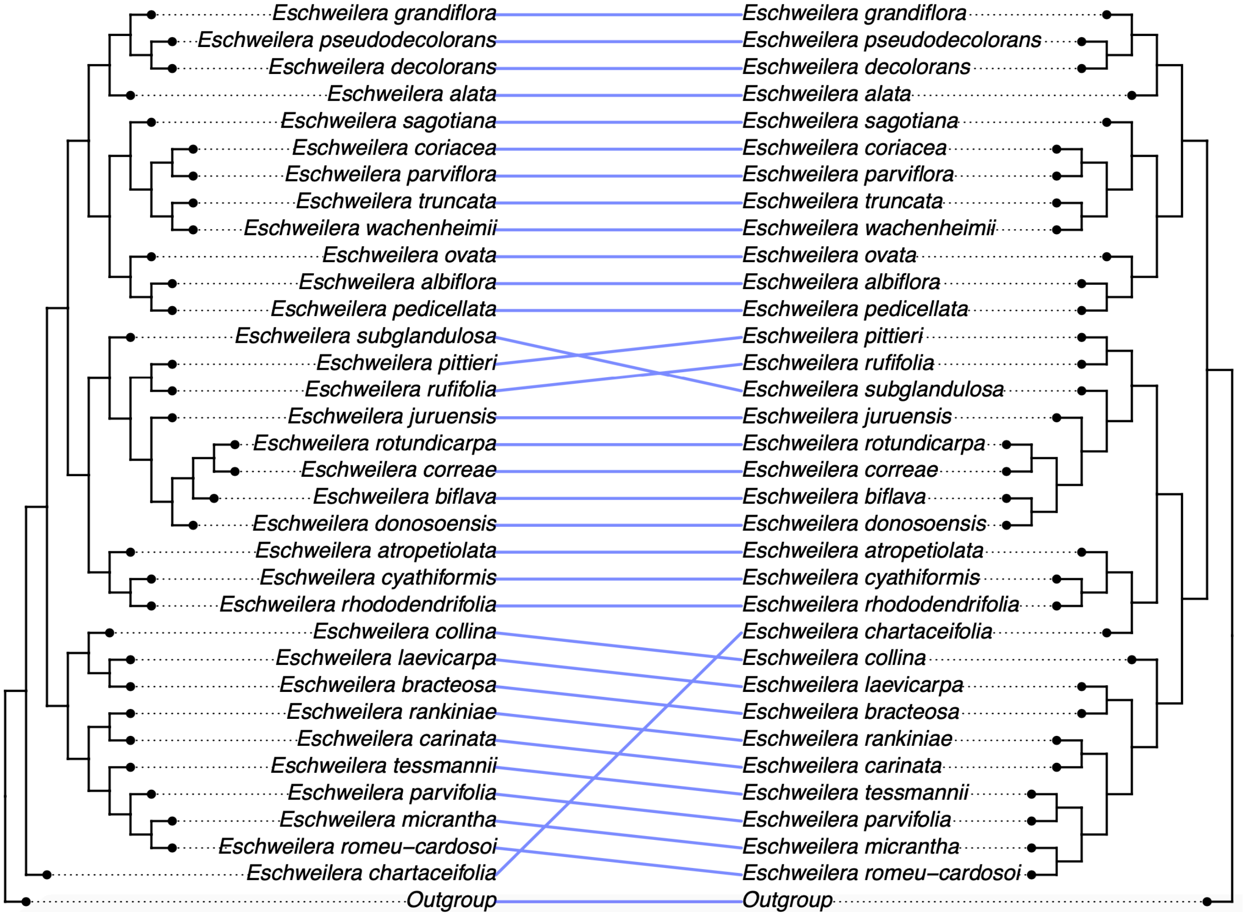


Figure B-8. Boxplots overlayed with dot plots showing the day of the year that collections of *Eschweilera coriacea* (n=35), *E. parviflora* (n=18), and *E. wachenheimii* (n=9) were made from Amazonas, Brazil for specimens housed at the New York Botanical Garden Herbarium (NY). Each dot represents a collection from a unique individual with open flowers or flower buds. The upper and lower limits of boxes represent the third and first quartiles respectively. Whiskers show the minimum and maximum values that occur within 1.5 times the length of the interquartile range. Points beyond the whiskers could be considered outliers. The darker midlines represent medians. On the y-axis, day 1 represents January 1st and day 300 represents October 27th in non-leap years.

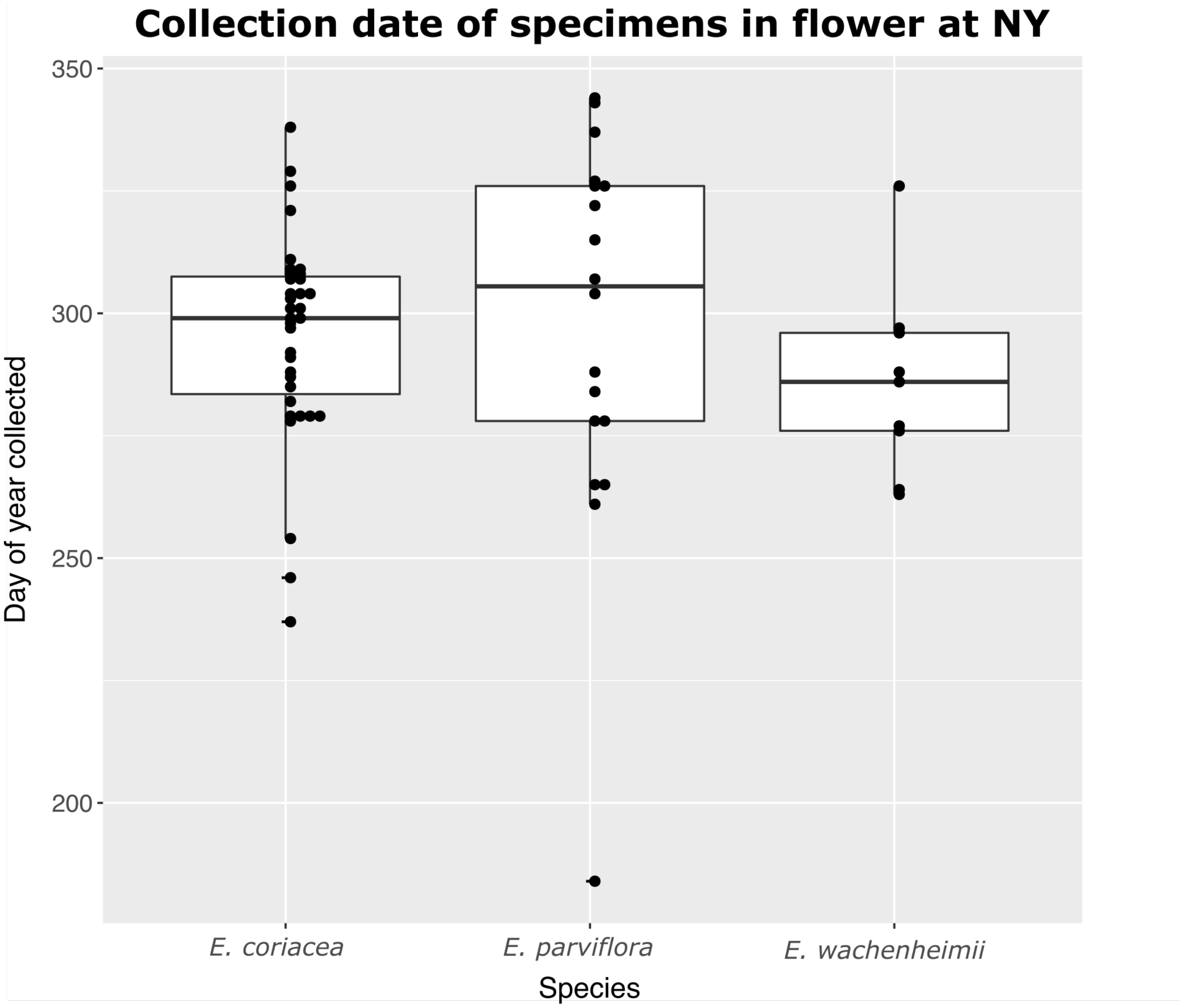


Table B-1. Voucher and accession information for samples used in the study. National Center for Biotechnology Information SRA accession numbers for all samples are listed, as are summarized results used to make redeterminations based on all available evidence. “Not applicable” is abbreviated as “n.a.”.

Available at: https://dx.doi.org/10.7302/4153

Table B-2. Summary statistics for all SNP datasets and estimated probability of the data for all *Structure* analyses for differing values of K. “Not applicable” is abbreviated as “n.a.”.

Available at: https://dx.doi.org/10.7302/4153

Table B-3. Summary of all rooted triplet tests conducted.

Available at: https://dx.doi.org/10.7302/4153

Table B-4. Results of tree searches and likelihood calculations for the Parvifolia phylogenies, ordered by increasing AIC score.

Available at: https://dx.doi.org/10.7302/4153

Table B-5. Morphological and ecological traits of species inferred to engage in admixture.

Available at: https://dx.doi.org/10.7302/4153

Table B-6. Information for specimens with flowers or flower buds collected in Amazonas, Brazil and housed at the New York Botanical Garden Herbarium.

Available at: https://dx.doi.org/10.7302/4153

Methods B-1. Paralog filtering and alignment.

We applied a tree-based approach to filtering the gene assemblies produced by HybPiper. While the probe set we used for target capture sequencing was meant to recover the same genes in all samples (i.e., orthologs), in cases where two or more similar sequences match a probe, both can be enriched during library preparation. The settings we used to assemble the data in HybPiper were such that when more than one contig is recovered at similar read coverage (less than 10x difference) for a given gene target and both assemblies were at least 85% the length of the reference sequence, the program returned whichever had the greatest percent identity to the reference. As an additional step to limit the inclusion of paralogs in the ortholog groups (i.e., orthogroups), we used *a priori* knowledge of the major clades of Lecythidaceae to split orthogroups, while retaining nearly all data for downstream analysis.

We first aligned exon and amino acid sequences for each of the 343 orthogroups recovered by HybPiper. All alignments were generated using MAFFT v7.271 and the option --maxiterate 1000. Amino acid sequences were aligned with the L-INS-i algorithm and nucleotide sequences were aligned using the less computationally intensive FFT-NS-i algorithm. A phylogeny for each was estimated using RAxML v8.2.11 with the GTRCAT model of evolution for nucleotide alignments and the PROTCATWAG model for amino acid alignments.

Then, we visually inspected orthogroup alignments for evidence of paralogy issues such as regions with many mis-matched bases, regions with lots of gaps, and instances of nearly identical sequences occurring in some taxa but not others in a biologically implausible way based on previous phylogenetic studies. In cases where visual inspection of an orthogroup alignment suggested paralogy issues, we found that phylogenies estimated from amino acid data tended to clearly reflect these issues. Many such amino acid phylogenies contained long branches (i.e. relatively many inferred substitutions per site), subtending "clades" which were extremely unlikely to occur due to biological processes, based on our understanding of the Lecythidaceae phylogeny. For example, the Neotropical Lecythidaceae (sometimes referred to as subfamily Lecythidoideae) has been strongly supported as a clade, and gene trees in which these species do not form a monophyletic group may be suspect. While phylogenies built using the corresponding exon and/or intron data usually also contained the same biologically dubious relationships, the branch lengths were more variable than those of the amino acid trees, possibly because there was a high proportion of gaps inferred for these nucleotide alignments. Therefore, we chose to examine individual amino acid phylogenies, to identify instances of potential paralog issues.
Amino acid trees without apparent paralog issues were used to determine the range amino acid substitutions per site inferred for the branch separating the Neotropical Lecythidaceae from the other members of the family. In nearly all cases where no issue was detected, the branch in question had a branch length less than 0.25 substitutions per site and this value was therefore used as the upper limit of expected branch lengths for genuinely orthologous sequences in amino acid trees. Using the methods of Yang & Smith (2014), we then cut any internal branches in the amino acid trees longer that 0.25 and retained any subtree with at least 10 taxa as a separate orthogroup. Any terminal branch in amino acid trees longer than 0.15 substitutions per site was also cut, to reduce paralogs occurring in any single sample. Following this procedure, there were 661 orthogroups. The corresponding nucleotide orthogroups (both exon and intron data) were then split to match the results of the amino acid orthogroup pruning.

Nucleotide sequences for each of the 661 resulting orthogroups were aligned separately with the L-INS-i algorithm in MAFFT.

Methods B-2. Genotyping and SNP dataset analyses.

Exon sequences from a single sample (i.e. *Eschweilera coriacea*; EscoL834) were used as a reference "genome" because we were able to recover sequence data for 343 loci for this sample and *E. coriacea* was the most extensively sampled species in our dataset. The custom script *make_reference_genome_from_exonerate_exons.py* was used to generate the reference from exon data using the *exonerate_results.fasta* file generated for this sample by HybPiper. For each target locus, exons were concatenated with a 400 "N" spacer between each to produce a single "pseudo-contig", meant to preserve linkage among exons of the same gene while reducing the possibility that read mapping errors could be caused by concatenating exon sequences in a non-biological way. Raw, untrimmed reads from target capture sequencing were concatenated with reads from whole genome shotgun sequencing of unenriched libraries. Then, SAM files were generated from raw reads for each of the 109 members of the Parvifolia clade (and one member of *Eschweilera* from outside the Parvifolia clade; *Eschweilera integrifolia*; sample EsinLA01) using the script *generate_sam_files_from_raw_reads_parvifolia.py*. Next, the custom script *GATK_sam_to_halplotypeCaller_parvifolia.py* was run for each sample in a Docker v18.09.7 container, which executed the following commands using The Genome Analysis Toolkit (GATK) v4.1.0.0 and Picard v2.18.25: 1) **samtools view** 2) **SortSam** 3) **MarkDuplicates** 4) **AddOrReplaceReadGroups** 5) **samtools index** 6) **HaplotypeCaller**. This resulted in a Genomic Variant Call Format (GVCF) file for each sample that was combined with the GATK command **CombineGVCFs**. Finally, variant calling was conducted using the command **GenotypeGVCFs** with the options "--include-non-variant-sites" and "--annotate-with-num-discovered-alleles true" and non-SNP variants were removed with the command **SelectVariants** and the options "--exclude-non-variants true", "--exclude-filtered true" and "--select-type-to-include SNP". This resulted in a Variant Call Format (VCF) file with SNP data for 110 samples that could later be subset to address specific questions regarding the population structure of species of the Parvifolia clade. The **SelectVariants** command was also used to exclude sample EsinLA01 (which is not a member of the Parvifolia clade) in order to confirm that its inclusion did not affect the total number of polymorphic sites identified, which was 148,310.

Subsets of samples were selected for analysis in *Structure* v2.3.4 based on the results of the preliminary phylogeny. For each taxon subset, variants were filtered using the programs plink (http://pngu.mgh.harvard.edu/purcell/plink/) and plink2 (www.cog-genomics.org/plink/2.0/) to obtain a subset of SNPs for that subset of individuals that were in approximate linkage equilibrium. This was accomplished using the following commands where X.plink.txt and X.pops.txt were files that specified the following respectively: 1) the samples to include in the subset and 2) *a priori* population membership based on previous morphological determination and the results of the preliminary phylogeny:

```
plink --vcf ../Genotypes_110_parvifolia_clade_EscoL834_ref_SNP_only.vcf --keep X.plink.txt --allow-extra-chr --make-bed

plink2 --allow-extra-chr --bfile plink --set-all-var-ids @_#_\$r_\$a --out prefilter --new-id-max-allele-len 286 --max-alleles 50 --make-bed

plink2 -bfile prefilter --indep-pairwise 50kb 1 0.0001 --allow-extra-chr -out LD_STEP

plink2 --allow-extra-chr --bfile prefilter --out final --new-id-max-allele-len 286 --max-alleles 50 --make-bed --extract LD_STEP.prune.in

plink --recode structure --allow-extra-chr -bfile final

python convert_plinkRecode_to_structure.py plink.recode.strct_in X.pops.txt
```

All *Structure* analyses were run for 1x10^{6} MCMC generations after 1x10^{5} generations of burnin. A full list of all parameters used in *Structure* runs is available in the Dryad repository submission accompanying this article. For each SNP dataset, *Structure* was run separately with the maximum number of populations (K) from 1 to 7, when we expected five or fewer clusters, or from 1 to 10 if the subset was expected to contain six or more clusters. Results of *Structure* analyses were formatted with the custom script *f2R.py* and visualized using custom R scripts.

The optimal K for each subset was determined by comparing the "Estimated Ln Prob of Data" (herein referred as the score) in the output files of each run with *a priori* taxonomic information. In most cases, the run with the best score matched or nearly matched our *a priori* expectation of the number of species based on morphological determinations and preliminary phylogenetic analysis. Exceptions to this occurred for taxa subsets where there were multiple species with only a single representative in the analysis and when there were multiple inferred clusters within one species. Though some of the species we included in our *Structure* analyses were represented by a single individual, we believe the results of these analyses are robust because 1) the loci targeted by our sequencing were selected *a priori* based on phylogenetic informativeness; 2) the results of these tests were concordant with our understanding of the Lecythidaceae phylogeny; 3) our SNP datasets averaged only 28.17% missing data and averaged 229

SNPs with no missing data for any individual; 4) results from runs with overlapping of individuals corroborate one another with respect to the ancestry of individual samples.

Methods B-3. Redetermination of individuals based on all available evidence.

Closely related species of Lecythidaceae are known to be difficult to identify, especially if reproductive material is unavailable and determinations must be based made on vegetative characters alone. In cases were the most recent morphological determination did not agree with the results of genetic analyses (i.e. the preliminary phylogeny, *Structure* analyses, and any relevant RT tests) all available evidence was used to determine the most likely species identity for the individual if it was shown to be a member of the Parvifolia clade in the preliminary phylogeny. For samples that fell outside the Parvifolia clade in preliminary phylogeny and were inferred to belong to a genus other than that of their most recent morphological determination, we assigned a genus-level redetermination (i.e. *Lecythis sp.*); such redeterminations had no effect on downstream analyses for this study, but were made in order to provide more accurate taxonomy for these samples in online databases and to facilitate their use in other studies. With one exception (i.e. *Eschweilera roseocalyx*, discussed below), a redetermination based on genetic evidence was made for individuals in the Parvifolia clade if a *Structure* analysis showed that the individual clustered most closely with individuals other than of its most recent morphological determination and these results were corroborated by phylogenetic analyses. Individuals were assigned a "*species affinis*" (i.e. aff.), designation to indicate alliance with a species other than that of their morphological determination if they met the following criteria: 1) The sample showed no significant evidence of admixture; 2) the sample had a species-level morphological determination and was not the only sample with that morphological determination; 3) phylogenetic analysis showed the sample was more closely related to another species than it was to other samples with its own morphological determination and 4) in phylogenetic analysis, the sample was inferred to diverge earlier than all other individuals of the species with which it was inferred to be most closely related, ignoring any individuals whose morphological and genetic determinations disagreed.

A recently described species, *Eschweilera roseocalyx*, was included our sampling (i.e. EscoL545). This species was described by Batista et al. (2017) based on a single individual found in a cloud forest in Panama's Chagres National Park. Our sampling included only one individual from Panama that was determined to be *E. coriacea* based on morphology; results from the preliminary phylogeny showed this individual was not a member of the Parvifolia clade and was likely mis-identified. Later phylogenomic analyses conducted in this study produced results that were compatible with an interpretation of *E. roseocalyx* as a geographically restricted ecotype of *E. coriacea*. Because of this compatibility and the absence of population-level data available elsewhere in the literature, we treat this individual as a member of *E. coriacea*. Species delimitation is outside the scope of the present work and we do not wish for our treatment of *E. roseocalyx* here to be interpreted as evidence against the taxonomic validity of any species. Subsequent studies may show this to be an example of peripatric speciation whereby the widespread *E. coriacea* has given rise to a reproductively distinct endemic species. However, given the available evidence and the focal questions the present work seeks to address, we feel that treating this individual as *E. coriacea* is more justifiable than treating it as a separate species in our results and discussion.

Methods B-4. Parvifolia phylogeny supermatrix construction and phylogeny estimation.

Orthogroup alignments used for the preliminary phylogenies were subset to include only members of the Parvifolia clade and five members of the Integrifolia clade as outgroups. Samples with evidence of recent admixture were excluded from this analysis, since admixture violates the assumptions of tree-based phylogenetic inference. We choose not to re-align orthogroups at this step, allowing sequences from taxa outside the Parvifolia clade to inform the final alignment of this dataset. This decision was due mainly to the fact that intron sequence recovery was highly variable among samples, which is to be expected because the probes used for the target capture sequencing protocol are meant to capture exons. The intronic sequences recovered are expected to be those adjacent to the targeted exon regions, captured because they are located close enough to the targeted exons to be enriched during library preparation, but outside the region for which the probes were specifically designed. This can result in greater variability in recovery during sequencing and assembly. The resulting alignments will likely have more missing data, especially if some intronic regions are only recovered in small number of taxa, which could increase alignment error. Poor alignment of sequences can severely impact model parameterization, likelihood calculations, and the inferred topology of phylogenies, and we therefore chose to include all available data in generating our orthogroup alignments.

To further reduce possible non-orthology in our estimation of the Parvifolia phylogeny, we employed a second tree-based filtering protocol to generate the final supermatrices. To accomplish this additional filtering, for each orthogroup (i.e. each of the 661 the resulted from the first round of trimming for the preliminary phylogeny), a tree was estimated separately for each intron and exon alignment using IQ-TREE with a GTR+Γ model of evolution after filtering out columns with less than 30% occupancy with the pxclsq command in phyx. For each unrooted tree, the average length and standard deviation of internal branch lengths and terminal branch lengths was calculated using the custom script *trim_trees_based_on_branch_distributions.py*. For this procedure, internal branch lengths were compared to other internal branches and terminal branch lengths were compared to other terminal branches. This allowed us, in an automated way, to determine branch lengths that were much longer than those of comparable branches within an orthogroup, which could be indicative of orthology issues. Any internal branch longer than average plus two standard deviations was cut using the methods of Yang & Smith (2014) and any terminal branches longer than average plus two standard deviations was also cut from the resulting subtrees. Any subtree without at least 27 tips representing members of the Parvifolia clade and five outgroups (approximately 25% occupancy) was excluded from further analysis. The resulting pruned trees were used to generate the final alignments by subsetting the orthogroup alignments to remove any sequences cut during the trimming procedure. Any sequence that would have contained more than 75% missing data in the final alignment was also excluded. Finally, sites with less than 30% occupancy were removed from each filtered alignment with the pxclsq command in phyx. Two supermatrices were produced with the pxcat command in phyx, one that included only exon data (668,353 aligned sites, 382 partitions, 29.74% missing data), and a second that included both exon and intron data (2,085,546 aligned sites, 765 partitions, 36.3% missing data).

For each supermatrix, 1) a maximum likelihood (ML) tree was estimated using RAxML v8.2.11 with a separate GTRCAT model of evolution specified for each partition with the -q option, 2) an ML tree was estimated using IQ-TREE v1.6.9 with a separate GTR+Γ model for each partition using the -q option and 3) a second ML tree was estimated with IQ-TREE and the same settings but allowing for partition-specific scaled evolutionary rates with the -spp option. To allow for direct comparisons, likelihoods and information criteria scores for each tree were recalculated with IQ-TREE with the following criteria 1) a tree topology fixed to that of the original result 2) a separate GTR+Γ model for each partition with and without partition-specific rates in separate analyses and 3) re-estimated branch lengths for the tree. The Akaike information criterion (AIC) score for each result was compared to determine the best scoring phylogeny for each supermatrix, which we considered as representing the best topological hypothesis for that dataset.

Notes B-1. Note on species of the Parvifolia clade at Reserve 1501.

There were five species known to occur in the 100-ha Lecythidaceae plot at Reserve 1501 that are now recognized as members of the Parvifolia clade, but that we did not include in our focal sampling either due to their rarity in the plot or past uncertainty in their phylogenetic placement. These non-focal species, all of which were represented by at least one individual in our overall sampling, were *E. carinata* S.A.Mori, *E. grandiflora* (Aubl.) Sandwith, *E. parviflora* (Aubl.) Miers, *E. romeu-cardosoi* S.A.Mori*,* and *E. tessmannii* R.Knuth.

Notes B-2. Notes on sampling at Reserve 1501 and prioritization of morphological intermediates.

We employed a mixture of planned and opportunistic sampling (i.e. adjusting sampling plans in the field and collecting obtainable specimens as they are discovered, rather than in a strictly randomized way) while making new collections at Reserve 1501. Various logistical challenges can make sampling from tropical trees difficult, even in established plots. For example, tree tags may fall off, making the target tree difficult or impossible to locate. Trees may have died since the last census, species may grow in patchy distributions so that sampling direct relatives may be a concern, and some trees may simply be too large or tall to safely collect a specimen. These logistical realities mean that opportunistically sampling can be a useful way to collect specimens in tropical forests.

In the vast majority of cases, we sampled trees based solely on the species name assigned during previous censuses of the plot in order to achieve a sampling rate of 4-8 individuals per focal species. We chose target trees to visit based on their previous species determination, that they were a minimum distance of 100m from any other sampled individual of their species, and that they were along a route that would facilitate collection of multiple samples during the day of field work. If, for whatever reason, a collection could not be safely obtained from a target tree, a substitute was chosen using the same criteria outlined above.

There were three trees we encountered in the Reserve 1501 plot that appeared to us to be possible hybrids between *Eschweilera coriacea* and another species. These trees displayed morphological traits, including branching architecture, that appeared to be intermediate between species. We intentionally prioritized these three samples for collection and genetic analysis. While one of these specimens did ultimately show genetic evidence of admixture between *E. coriacea* and *E. wachenheimii*, the other two showed no evidence of admixture. However, a fourth sample from Reserve 1501, thought to be *E. coriacea* in the field, did show strong evidence of admixture in later analyses. Outside of Reserve 1501, neither individual ultimately found to be *E. parviflora* × *wachenheimii* based on genetic evidence was suspected of being admixed prior to genetic analysis. Thus, it should be noted that most individuals with genetic evidence of admixture in our study were not suspected to be admixed based on morphology.

.

Appendix C

Supplementary Tables and Notes for Chapter IV

Table C-1. Accession information for samples used in this study.

Available at: https://dx.doi.org/10.7302/4153

Table C-2. Recognized genera of Primulaceae, synonymy, and representation in this study.

In final sampling?	Subfamily	Genus	Authority	Accepted by POWO?	Synonym according to POWO
yes	Maes.	*Maesa*	Forssk.	yes	
yes	Myrsin.	*Aegiceras*	Gaertn.	yes	
yes, *Afrardisia buesgenii* Gilg & G.Schellenb.	Myrsin.	*Afrardisia*	Mez	no	(=Ardisia Sw.)
yes	Myrsin.	*Amblyanthopsis*	Mez	yes	
no, *Amblyanthus glandulosus* (Roxb.) A.DC., excluded due to QC	Myrsin.	*Amblyanthus*	A.DC.	yes	
yes, *Anagallis minima* (L.) E.H.L.Krause	Myrsin.	*Anagallis*	L.	no	(=Lysimachia Tourn. ex L.)
yes	Myrsin.	*Antistrophe*	A.DC.	yes	
yes	Myrsin.	*Ardisia*	Sw.	yes	
yes	Myrsin.	*Ardisiandra*	Hook.f.	yes	
yes	Myrsin.	*Badula*	Juss.	yes	
yes	Myrsin.	*Conandrium*	(K.Schum.) Mez	yes	
yes	Myrsin.	*Coris*	L.	yes	
no, *Ctenardisia amplifolia*, excluded due to QC	Myrsin.	*Ctenardisia*	Ducke	yes	
yes	Myrsin.	*Cybianthus*	Mart.	yes	
yes	Myrsin.	*Cyclamen*	L.	yes	
yes	Myrsin.	*Discocalyx*	(A.DC.) Mez	yes	
yes	Myrsin.	*Elingamita*	G.T.S.Baylis	yes	
yes	Myrsin.	*Embelia*	Burm.f.	yes	
yes	Myrsin.	*Emblemantha*	B.C.Stone	yes	

yes	Myrsin.	*Fittingia*	Mez	yes	
yes	Myrsin.	*Geissanthus*	Hook.f.	yes	
yes, *Gentlea micranthera* (Donn.Sm.) Lundell	Myrsin.	*Gentlea*	Lundell	no	(=Ardisia Sw.)
no	Myrsin.	*Grammadenia*	Bentham	no	(=Cybianthus Mart.)
no, *Grenacheria buxifolia*, excluded due to QC	Myrsin.	*Grenacheria*	Mez	no	(=Embelia Burm.f.)
yes	Myrsin.	*Heberdenia*	Banks ex A.DC.	yes	
yes	Myrsin.	*Hymenandra*	A.DC. ex Spach	yes	
yes	Myrsin.	*Labisia*	Lindl.	yes	
yes	Myrsin.	*Loheria*	Merr.	yes	
yes	Myrsin.	*Lysimachia*	Tourn. ex L.	yes	
no	Myrsin.	*Mangenotiella*	M.Schmid	yes	
no	Myrsin.	*Microconomorpha*	(Mez) Lundell	no	(=Cybianthus Mart.)
yes	Myrsin.	*Monoporus*	A.DC.	yes	
yes	Myrsin.	*Myrsine*	L.	yes	
yes	Myrsin.	*Oncostemum*	A.Juss.	yes	
no, not recognized, see main text	Myrsin.	*Paralysimachia*	F.Du, J.Wang & S.Y.Yang	no, not recognized here. See main text.	na
yes	Myrsin.	*Parathesis*	Hook.f.	yes	
yes	Myrsin.	*Pleiomeris*	A.DC.	yes	
yes, *Rapanea guyanensis* (Aubl.) Kuntze	Myrsin.	*Rapanea*	Aublet	no	(=Myrsine L.)
yes	Myrsin.	*Sadiria*	Mez	yes	
no	Myrsin.	*Solonia*	Urb.	yes	
yes	Myrsin.	*Stimpsonia*	C.Wright ex A.Gray	yes	
yes	Myrsin.	*Stylogyne*	A.DC.	yes	
yes, *Synardisia venosa* (Mast. ex Donn.Sm.) Lundell	Myrsin.	*Synardisia*	(Mez) Lundell	no	(=Ardisia Sw.)
yes	Myrsin.	*Systellantha*	B.C.Stone	yes	
yes	Myrsin.	*Tapeinosperma*	Hook.f.	yes	
no, *Tetrardisia porosa* (C.B.Clarke) Furtado, excluded due to QC	Myrsin.	*Tetrardisia*	Mez	no	(=Ardisia Sw.)
yes, *Trientalis europaea* (L.) U.Manns & Anderb	Myrsin.	*Trientalis*	L.	no	(=Lysimachia Tourn. ex L.)
no	Myrsin.	*Vegaea*	Urb.	yes	
yes	Myrsin.	*Wallenia*	Sw.	yes	
no	Myrsin.	*Yunckeria*	Lundell	no	(=Ctenardisia Ducke)
yes	Primul.	*Androsace*	L.	yes	

yes	Primul.	*Bryocarpum*	Hook.f. & Thomson	yes	
yes, *Cortusa brotheri* (R.Knuth) Losinsk.	Primul.	*Cortusa*	L.	no	(=Primula L.)
yes	Primul.	*Dionysia*	Fenzl	yes	
no, not recognized	Primul.	***Evotrochis***	Raf.	no, not recognized here. See main text.	(=Primula L.)
yes	Primul.	*Hottonia*	Boerh. ex L.	yes	
yes	Primul.	*Kaufmannia*	Regel	yes	
yes	Primul.	*Omphalogramma*	Franch.	yes	
yes	Primul.	*Pomatosace*	Maxim.	yes	
yes	Primul.	*Primula*	L.	yes	
yes	Primul.	*Soldanella*	L.	yes	
yes, *Vitaliana primuliflora*	Primul.	*Vitaliana*	Sesl.	no	(=Androsace L.)
yes	Theophrast.	*Bonellia*	Bertero ex Colla	yes	
yes	Theophrast.	*Clavija*	Ruiz & Pav.	yes	
yes	Theophrast.	*Deherainia*	Decne.	yes	
yes	Theophrast.	*Jacquinia*	L.	yes	
yes	Theophrast.	*Neomezia*	Votsch	yes	
yes	Theophrast.	*Samolus*	L.	yes	
yes	Theophrast.	*Theophrasta*	L.	yes	
no	Theophrast.	*Votschia*	B.Ståhl	yes	

Notes C-1. Additional descriptions of conflict between species trees.

Comparison of Ericales1-exon-ML & Ericales1-exon-ASTRAL

Within Primulaceae, there were 39 nodes that conflicted between Ericales1-exon-ML and Ericales1-exon-ASTRAL and 109 (i.e. not counting the root) that agreed. Outside Myrsinoideae, only two nodes differed. In Ericales1-exon-ASTRAL, *Samolus* was not sister to the other Theophrastoideae, but was instead sister to Myrsinoideae + Primuloideae, with the other Theophrastoideae sister to that clade. The other difference was that in Ericlaes1-exon-ASTRAL, *P. frigida* + *P. suffrutescens and Kaufmannia semenovii* + *Primula matthioli* were sister, whereas the two were successively sister to all other *Primula* + *Dionysia* in Ericales1-exon-ML.

Comparison of Ericales1-exon-ML & Prim2-exon-ML

The topologies within Primulaceae of the trees Ericales1-exon-ML and Prim2-exon-ML were very similar as might be expected since they are based on the same underlying exon dataset. Relationships among subfamilies was the same. Relationships among genera were identical in both trees, except that in Ericales1-exon-ML, *Ardisiandra* was sister to *Cyclamen* + *Lysimachia*, while in Prim2-exon-ML the position of *Ardisiandra* was one node back, putting it along the backbone of Myrsinoideae, such that *Ardisiandra*, *Coris*, and *Stimpsonia* were successively sister to all other Myrsinoideae. Regarding conflict within genera, In the Ericales1-exon-ML tree, *Lysimachia nummularia* and *L. clethroides* were successively sister to a clade of *L. vulgas* + *L. quadrifolia* + *L. terrestris*. In the Prim2-exon-ML, the positions of *L. nummularia* and *L. clethroides* were transposed. Within *Oncostemum*, four nodes differed between the two datasets, and four were concordant. Overall, 142 of the 148 internal nodes within Primulaceae (i.e. not counting the root) were concordant between the two trees.

Comparison of Ericales1-exon-ML and Prim3-intron-ML

In the Ericales1-exon-ML, *Cybianthus perseoides* and *C. pastensis* were successively sister to the rest of *Cybianthus*, while in the Prim3-intron-ML, those two formed a clade that was sister to the rest of *Cybianthus*. All other relationships outside the Old World Ardisioids were identical between the Ericales1-exon-ML and the Prim3-intron-ML. Within the Old World Ardisioids, 22 nodes were concordant and 17 nodes conflicted between the Ericales1-exon-ML and the Prim3-intron-ML. The Prim3-intron-ML tree placed *Ardisiandra* sister to *Cylamen* + *Lysimachia* with ultrafast bootstrap support of 85%, while this relationship received 87% ultrafast bootstrap support in Ericales1-exon-ML, indicating that this relationship is uncertain in both trees.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anderberg, A.A., Rydin, C., Källersjö, M., 2002. Phylogenetic relationships in the order Ericales sl: analyses of molecular data from five genes from the plastid and mitochondrial genomes. American Journal of Botany 89, 677–687.

Anderberg, A.A., Ståhl, B., 1995. Phylogenetic interrelationships in the order Primulales, with special emphasis on the family circumscriptions. Canadian Journal of Botany 73, 1699–1730. https://doi.org/10.1139/b95-184

Anderberg, A.A., Ståhl, B., Källersjö, M., 2000. Maesaceae, a new primuloid family in the order Ericales s.l. Taxon 49, 183–187. https://doi.org/10.2307/1223834

Anderberg, A.A., Ståhl, B., Källersjö, M., 1998. Phylogenetic relationships in the Primulales inferred from rbcL sequence data. Plant Systematics and Evolution 211, 93–102. https://doi.org/10.1007/BF00984914

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data [WWW Document]. URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc

APG I, 1998. An ordinal classification for the families of flowering plants. Annals of the Missouri Botanical Garden 85, 531–553. https://doi.org/10.2307/2992015

APG II, 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Botanical Journal of the Linnean Society 141, 399–436. https://doi.org/10.1046/j.1095-8339.2003.t01-1-00158.x

APG III, 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society 161, 105–121. https://doi.org/10.1111/j.1095-8339.2009.00996.x

APG IV, 2016. 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, 1–20. https://doi.org/10.1111/boj.12385

Appelhans, M.S., Paetzold, C., Wood, K.R., Wagner, W.L., 2020. RADseq resolves the phylogeny of Hawaiian *Myrsine* (Primulaceae) and provides evidence for hybridization. Journal of Systematics and Evolution. https://doi.org/10.1111/jse.12668

Arias, T., Beilstein, M.A., Tang, M., McKain, M.R., Pires, J.C., 2014. Diversification times among *Brassica* (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. American Journal of Botany 101, 86–91.

Ashton, P.S., 1969. Speciation among tropical forest trees: some deductions in the light of recent evidence. Biological Journal of the Linnean Society 1, 155–196. https://doi.org/10.1111/j.1095-8312.1969.tb01818.x

Baker, W.J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., Botigué, L.R., Brewer, G., Carruthers, T., Clarkson, J.J., Cook, J., Cowan, R.S., Dodsworth, S., Epitawalage, N., Françoso, E., Gallego, B., Johnson, M.G., Kim, J.T., Leempoel, K., Maurin, O., Mcginnie, C., Pokorny, L., Roy, S., Stone, M., Toledo, E., Wickett, N.J., Zuntini, A.R., Eiserhardt, W.L., Kersey, P.J., Leitch, I.J., Forest, F., 2021. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. Systematic Biology. https://doi.org/10.1093/sysbio/syab035

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Batista, J.E., Mori, S.A., Harrison, J.S., 2017. New species of *Eschweilera* and a first record of *Cariniana* (Lecythidaceae) from Panama. Phytoneuron 2017–62, 1–16.

Baylis, G.T.S., 1951. *Elingamita* (Myrsinaceae) a new monotypic genus from West Island, Three Kings. Records of the Auckland Institute and Museum 4, 99–102.

Berchtold, B.V., Presl, J.S., 1820. O přirozenosti rostlin aneb rostlinář obsahugjej i gednanj o žiwobytj rostlinném pro sebe az ohledu giných žiwotů podlé stawu nyněgssjho znánj. KW Enders.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bone, R.E., Strijk, J.S., Fritsch, P.W., Buerki, S., Strasberg, D., Thebaud, C., Hodkinson, T.R., 2012. Phylogenetic inference of *Badula* (Primulaceae), a rare and threatened genus endemic to the Mascarene Archipelago. Botanical Journal of the Linnean Society 169, 284–296.

Boucher, F.C., Zimmermann, N.E., Conti, E., 2016. Allopatric speciation with little niche divergence is common among alpine Primulaceae. Journal of Biogeography 43, 591–602. https://doi.org/10.1111/jbi.12652

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLOS Computational Biology 10, e1003537. https://doi.org/10.1371/journal.pcbi.1003537

Bremer, B., Bremer, Ka., Heidari, N., Erixon, P., Olmstead, R.G., Anderberg, A.A., Källersjö, M., Barkhordarian, E., 2002. Phylogenetics of asterids based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. Molecular Phylogenetics and Evolution 24, 274–301.

Brewer, G.E., Clarkson, J.J., Maurin, O., Zuntini, A.R., Barber, V., Bellot, S., Biggs, N., Cowan, R.S., Davies, N.M.J., Dodsworth, S., Edwards, S.L., Eiserhardt, W.L., Epitawalage, N., Frisby, S., Grall, A., Kersey, P.J., Pokorny, L., Leitch, I.J., Forest, F., Baker, W.J., 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. Frontiers in Plant Science 10, 1102. https://doi.org/10.3389/fpls.2019.01102

Brown, J.M., Thomson, R.C., 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Systematic Biology 66, 517–530.

Brown, J.W., Walker, J.F., Smith, S.A., 2017a. Phyx: phylogenetic tools for unix. Bioinformatics 33, 1886–1888. https://doi.org/10.1093/bioinformatics/btx063

Brown, J.W., Wang, N., Smith, S.A., 2017b. The development of scientific consensus: Analyzing conflict and concordance among avian phylogenies. Molecular Phylogenetics and Evolution 116, 69–77.

Bryant, D., Hahn, M.W., 2020. The concatenation question. Phylogenetics in the Genomic Era 3–4. No commercial publisher | Authors open access book, 3.4:1--3.4:23.

Caesalpinus, A., 1583. De plantis libri. Giorgio Marescotti, Florence.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. https://doi.org/10.1186/1471-2105-10-421

Cannon, C.H., Lerdau, M., 2015. Variable mating behaviors and the maintenance of tropical biodiversity. Frontiers in Genetics 6, 183. https://doi.org/10.3389/fgene.2015.00183

Cantino, P.D., de Queiroz, K., 2020. International Code of Phylogenetic Nomenclature (PhyloCode): Version 6. CRC Press.

Caris, P.L., Smets, E.F., 2004. A floral ontogenetic study on the sister group relationship between the genus *Samolus* (Primulaceae) and the Theophrastaceae. American Journal of Botany 91, 627–643.

Caron, H., Molino, J.-F., Sabatier, D., Léger, P., Chaumeil, P., Scotti-Saintagne, C., Frigério, J.-M., Scotti, I., Franc, A., Petit, R.J., 2019. Chloroplast DNA variation in a hyperdiverse tropical tree community. Ecology and Evolution 9, 4897–4905. https://doi.org/10.1002/ece3.5096

Cavender-Bares, J., 2019. Diversification, adaptation, and community assembly of the American oaks (*Quercus*), a model clade for integrating ecology and evolution. New Phytologist 221, 669–692. https://doi.org/10.1111/nph.15450

Chase, M.W., Soltis, D.E., Olmstead, R.G., Morgan, D., Les, D.H., Mishler, B.D., Duvall, M.R., Price, R.A., Hills, H.G., Qiu, Y.-L., 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Annals of the Missouri Botanical Garden 528–580.

Chen, J.-D., Zheng, C., Ma, J.-Q., Jiang, C.-K., Ercisli, S., Yao, M.-Z., Chen, L., 2020. The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. Horticulture Research 7, 63. https://doi.org/10.1038/s41438-020-0288-2

Chernomor, O., von Haeseler, A., Minh, B.Q., 2016. Terrace aware data structure for phylogenomic inference from supermatrices. Systematic Biology 65, 997–1008. https://doi.org/10.1093/sysbio/syw037

Choi, J.Y., Purugganan, M., Stacy, E.A., 2020. Divergent selection and primary gene flow shape incipient speciation of a riparian tree on Hawaii Island. Molecular Biology and Evolution 37, 695–710. https://doi.org/10.1093/molbev/msz259

Christe, C., Boluda, C.G., Koubínová, D., Gautier, L., Naciri, Y., 2021. New genetic markers for Sapotaceae phylogenomics: more than 600 nuclear genes applicable from family to population levels. Molecular Phylogenetics and Evolution 160, 107123. https://doi.org/10.1016/j.ympev.2021.107123

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Couvreur, T.L.P., Helmstetter, A.J., Koenen, E.J.M., Bethune, K., Brandão, R.D., Little, S.A., Sauquet, H., Erkens, R.H.J., 2019. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. Frontiers in Plant Science 9, 1941. https://doi.org/10.3389/fpls.2018.01941

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V., Udall, J., 2012. Targeted enrichment strategies for next-generation plant biology. American Journal of Botany 99, 291–311. https://doi.org/10.3732/ajb.1100356

Cronquist, A., 1981. An integrated system of classification of flowering plants. Columbia University Press.

Darwin, C., 1859. On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. John Murray, London.

Darwin, C., Wallace, A., 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. Zoological Journal of the Linnean Society 3, 45–62. https://doi.org/10.1111/j.1096-3642.1858.tb02500.x

Davis, C.C., Xi, Z., 2015. Horizontal gene transfer in parasitic plants. Current Opinion in Plant Biology 26, 14–19. https://doi.org/10.1016/j.pbi.2015.05.008

De Queiroz, K., 2007. Species concepts and species delimitation. Systematic Biology 56, 879–886. https://doi.org/10.1080/10635150701701083

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43, 491–498. https://doi.org/10.1038/ng.806

Dick, C.W., Heuertz, M., 2008. The complex biogeographic history of a widespread tropical tree species. Evolution 62, 2760–2774. https://doi.org/10.1111/j.1558-5646.2008.00506.x

Dick, C.W., Pennington, R.T., 2019. History and geography of Neotropical tree diversity. Annual Review of Ecology, Evolution, and Systematics 50, 279–301. https://doi.org/10.1146/annurev-ecolsys-110617-062314

Doyle, J.J., 2022. Defining coalescent genes: theory meets practice in organelle phylogenomics. Systematic Biology 71, 476–489. https://doi.org/10.1093/sysbio/syab053

Drinkell, C., Utteridge, T.M.A., 2015. A revision of the genus *Systellantha* B. C. Stone. Studies in Malaysian Myrsinaceae IV. Kew Bulletin 70, 50. https://doi.org/10.1007/s12225-015-9603-8

Du, F., Juan, W., Yang, S., 2016. *Paralysimachia* F. Du, J. Wang et S. S. Yang, a new genus of Primulaceae from Yunnan. Journal of Southwest Forestry College 36, 91–94.

Dubéarnès, A., Julius, A., Utteridge, T.M., 2015. A synopsis of the genus *Embelia* in Peninsular Malaysia and Singapore. Studies in Malaysian Myrsinaceae III. Kew Bulletin 70, 25.

Eaton, D.A.R., Hipp, A.L., González-Rodríguez, A., Cavender-Bares, J., 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution 69, 2587–2601. https://doi.org/10.1111/evo.12758

Eaton, D.A.R., Ree, R.H., 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Systematic Biology 62, 689–706. https://doi.org/10.1093/sysbio/syt032

Eisen, J.A., 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research 8, 163–167.

Felsenstein, J., 1982. Numerical methods for inferring evolutionary trees. The Quarterly Review of Biology 57, 379–404.

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Biology 22, 240–249. https://doi.org/10.1093/sysbio/22.3.240

Firat, M., Lidén, M., 2021. The genus *Evotrochis* (Primulaceae) resurrected. Acta Biologica Turcica 34, 161–168.

Friis, E.M., Crane, P.R., Pedersen, K.R., 2021. Early flowers of primuloid Ericales from the Late Cretaceous of Portugal and their ecological and phytogeographic implications. Fossil Imprint 77, 214–230.

Friis, E.M., Pedersen, K.R., Crane, P.R., 2010. Cretaceous diversification of angiosperms in the western part of the Iberian Peninsula. Review of Palaeobotany and Palynology 162, 341–361. https://doi.org/10.1016/j.revpalbo.2009.11.009

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.

Gabaldón, T., 2008. Large-scale assignment of orthology: back to phylogenetics? Genome Biology 9, 1–6.

Geuten, K., Smets, E., Schols, P., Yuan, Y.-M., Janssens, S., Küpfer, P., Pyck, N., 2004. Conflicting phylogenies of balsaminoid families and the polytomy in Ericales: combining data in a Bayesian framework. Molecular Phylogenetics and Evolution 31, 711–729.

Gitzendanner, M.A., Soltis, P.S., Wong, G.K., Ruhfel, B.R., Soltis, D.E., 2018. Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. American Journal of Botany 105, 291–301.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nature Biotechnology 29, 644–652.

Graham, C.F., Glenn, T.C., McArthur, A.G., Boreham, D.R., Kieran, T., Lance, S., Manzon, R.G., Martino, J.A., Pierson, T., Rogers, S.M., Wilson, J.Y., Somers, C.M., 2015. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). Molecular Ecology Resources 15, 1304–1315. https://doi.org/10.1111/1755-0998.12404

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., Pääbo, S., 2010. A Draft Sequence of the Neandertal Genome. Science 328, 710. https://doi.org/10.1126/science.1188021

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8, 1494–1512.

Hahn, M.W., 2019. Molecular population genetics. Oxford University Press, New York,NY.

Hanlon, V.C., Otto, S.P., Aitken, S.N., 2019. Somatic mutations substantially increase the per-generation mutation rate in the conifer *Picea sitchensis*. Evolution Letters 3, 348–358.

Hao, G., Yuan, Y.-M., Hu, C.-M., Ge, X.-J., Zhao, N.-X., 2004. Molecular phylogeny of *Lysimachia* (Myrsinaceae) based on chloroplast trnL–F and nuclear ribosomal ITS sequences. Molecular Phylogenetics and Evolution 31, 323–339. https://doi.org/10.1016/S1055-7903(03)00286-0

Hardin, J.W., 1975. Hybridization and introgression in *Quercus alba*. Journal of the Arnold Arboretum 56, 336–363.

He, Y., Kueffer, C., Shi, P., Zhang, X., Du, M., Yan, W., Sun, W., 2014. Variation of biomass and morphology of the cushion plant *Androsace tapete* along an elevational gradient in the Tibetan Plateau. Plant Species Biology 29, E64–E71.

Hedwall, P., Brunet, J., Nordin, A., Bergh, J., 2013. Changes in the abundance of keystone forest floor species in response to changes of forest structure. Journal of Vegetation Science 24, 296–306.

Heenan, P., 2000. Dioecism in *Elingamita johnsonii* (Myrsinaceae). New Zealand Journal of Botany 38, 569–574.

Hennig, W., 1966. Phylogenetic Systematics. Univ. Illinois Press. Translated by D.D. Davis and R. Zangerl., Urbana, IL.

Hennig, W., 1950. Grundzuge einer theorie der phylogenetischen systematik. Deutscher Zentralverlag, Berlin.

Heuertz, M., Caron, H., Scotti-Saintagne, C., Pétronelli, P., Engel, J., Tysklind, N., Miloudi, S., Gaiotto, F.A., Chave, J., Molino, J.-F., 2020. The hyperdominant tropical tree *Eschweilera coriacea* (Lecythidaceae) shows higher genetic heterogeneity than sympatric *Eschweilera* species in French Guiana. Plant Ecology and Evolution 153, 67–81.

Hipp, A.L., Manos, P.S., Hahn, M., Avishai, M., Bodénès, C., Cavender-Bares, J., Crowl, A.A., Deng, M., Denk, T., Fitz-Gibbon, S., Gailing, O., González-Elizondo, M.S., González-Rodríguez, A., Grimm, G.W., Jiang, X.-L., Kremer, A., Lesur, I., McVay, J.D., Plomion, C., Rodríguez-Correa, H., Schulze, E.-D., Simeone, M.C., Sork, V.L., Valencia-Avalos, S., 2020. Genomic landscape of the global oak phylogeny. New Phytologist 226, 1198–1212. https://doi.org/10.1111/nph.16162

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Huang, S., Ding, J., Deng, D., Tang, W., Sun, Honghe, Liu, D., Zhang, L., Niu, X., Zhang, X., Meng, M., Yu, J., Liu, J., Han, Y., Shi, W., Zhang, D., Cao, S., Wei, Z., Cui, Y., Xia, Y., Zeng, H., Bao, K., Lin, L., Min, Y., Zhang, H., Miao, M., Tang, X., Zhu, Y., Sui, Y., Li, G., Sun, Hanju, Yue, J., Sun, J., Liu, F., Zhou, L., Lei, L., Zheng, X., Liu, M., Huang, L., Song, J., Xu, C., Li, J., Ye, K., Zhong, S., Lu, B.-R., He, G., Xiao, F., Wang, H.-L., Zheng, H., Fei, Z., Liu, Y., 2013. Draft genome of the kiwifruit *Actinidia chinensis*. Nature Communications 4, 2640. https://doi.org/10.1038/ncomms3640

Huang, Y.-Y., Mori, S.A., Kelly, L.M., 2015. Toward a phylogenetic-based generic classification of Neotropical Lecythidaceae—I. Status of *Bertholletia*, *Corythophora*, *Eschweilera* and *Lecythis*. Phytotaxa 203, 23. doi: 10.11646/phytotaxa.203.2.1.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23, 254–267. https://doi.org/10.1093/molbev/msj030

Huson, D.H., Klöpper, T., Lockhart, P.J., Steel, M.A., 2005. Reconstruction of reticulate networks from gene trees, in: Miyano, S., Mesirov, J., Kasif, S., Istrail, S., Pevzner, P.A., Waterman, M. (Eds.), Research in Computational Molecular Biology. Springer Berlin Heidelberg, Berlin, Heidelberg, 233–249.

IPNI (2021). International Plant Names Index. The Royal Botanic Gardens, Kew, Harvard University Herbaria & Libraries and Australian National Botanic Gardens. Published on the Internet; https://www.ipni.org/ [Accessed 21 January 2022].

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf, M., Ruzicka, D.R., Wafula, E., Wickett, N.J., 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biology 13, 1–14.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473, 97–100.

Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J.C., Wickett, N.J., 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Applications in Plant Sciences 4, 1600016. https://doi.org/10.3732/apps.1600016

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T., Leebens-Mack, J.H., Leitch, I.J., Maurin, O., Soltis, D.E., Soltis, P.S., Wong, G.K., Baker, W.J., Wickett, N.J., 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Systematic Biology 68, 594–606. https://doi.org/10.1093/sysbio/syy086

Judd, W.S., Campbell, C.S., Kellogg, E.A., Stevens, P.F., Donoghue, M.J., 2015. Plant Systematics: A Phylogenetic Approach, 4th ed. Sinauer Associates, Sunderland, MA.

Judd, W.S., Sanders, R.W., Donoghue, M.J., 1994. Angiosperm family pairs: preliminary phylogenetic analyses. Harvard Papers in Botany 1, 1–51.

Julius, A., Gutiérrez-Ortega, J.S., Sabran, S., Tagane, S., Naiki, A., Darnaedi, D., Aung, M., Dang, S., Nguyen, N., Binh, H., Watano, Y., Utteridge, T., Kajita, T., 2021. Phylogenetic relationship of tropical Asian *Ardisia* and relatives (Primulaceae) shows non-monophyly of recognized genera and subgenera. Journal of Japanese Botany 93, 149–165.

Julius, A., Utteridge, T.M.A., 2012. Revision of *Ardisia* subgenus *Bladhia* in Peninsular Malaysia; studies in Malaysian Myrsinaceae I. Kew Bulletin 67, 379–388. https://doi.org/10.1007/s12225-012-9374-4

Källersjö, M., Bergqvist, G., Anderberg, A.A., 2000. Generic realignment in primuloid families of the Ericales s.l.: a phylogenetic analysis based on DNA sequences from three chloroplast genes and morphology. American Journal of Botany 87, 1325–1341. https://doi.org/10.2307/2656725

Källersjö, M., Farris, J.S., Chase, M.W., Bremer, B., Fay, M.F., Humphries, C.J., Petersen, G., Seberg, O., Bremer, K., 1998. Simultaneous parsimony jackknife analysis of 2538rbcL DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. Plant Systematics and Evolution 213, 259–287. https://doi.org/10.1007/BF00985205

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30, 3059–3066. https://doi.org/10.1093/nar/gkf436

Katoh, K., Standley, D.M., 2013. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30, 772–780. https://doi.org/10.1093/molbev/mst010

Kearns, A.M., Restani, M., Szabo, I., Schrøder-Nielsen, A., Kim, J.A., Richardson, H.M., Marzluff, J.M., Fleischer, R.C., Johnsen, A., Omland, K.E., 2018. Genomic evidence of speciation reversal in ravens. Nature Communications 9, 906. https://doi.org/10.1038/s41467-018-03294-w

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A., 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35, 4453–4455.

Kremer, A., Hipp, A.L., 2020. Oaks: an evolutionary success story. New Phytologist 226, 987–1011. https://doi.org/10.1111/nph.16274

Larson, D.A., Vargas, O.M., Vicentini, A., Dick, C.W., 2021. Admixture may be extensive among hyperdominant Amazon rainforest tree species. New Phytologist 232, 2520–2534.

Larson, D.A., Walker, J.F., Vargas, O.M., Smith, S.A., 2020. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. American Journal of Botany 107, 773–789. https://doi.org/10.1002/ajb2.1469

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S.A., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J., DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B., Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M., Barrett, M.D., Baucom, R.S., Beerling, D.J., Benstein, R.M., Biffin, E., Brockington, S.F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L., Cannon, S.B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N.D., Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., dePamphilis, C.W., Der, J.P., Determann, R., Dickson, R.C., Di Stilio, V.S., Ellis, S., Fast, E., Feja, N., Field, K.J., Filatov, D.A., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gâteblé, G., Godden, G.T., Goh, F. (Qi Y., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K., Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Joyce, B., Kapralov, M.V., Kazamia, E., Kellogg, E.A., Koch, M.A., Von Konrat, M., Könyves, K., Kutchan, T.M., Lam, V., Larsson, A., Leitch, A.R., Lentz, R., Li, F.-W., Lowe, A.J., Ludwig, M., Manos, P.S., Mavrodiev, E., McCormick, M.K., McKain, M., McLellan, T., McNeal, J.R., Miller, R.E., Nelson, M.N., Peng, Y., Ralph, P., Real, D., Riggins, C.W., Ruhsam, M., Sage, R.F., Sakai, A.K., Scascitella, M., Schilling, E.E., Schlösser, E.-M., Sederoff, H., Servick, S., Sessa, E.B., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C., Smith, A.G., Smithson, A., Stewart, C.N., Stinchcombe, J.R., Szövényi, P., Tate, J.A., Tiebel, H., Trapnell, D., Villegente, M., Wang, C.-N., Weller, S.G., Wenzel, M., Weststrand, S., Westwood, J.H., Whigham, D.F., Wu, S., Wulff, A.S., Yang, Y., Zhu, D., Zhuang, C., Zuidof, J., Chase, M.W., Pires, J.C., Rothfels, C.J., Yu, J., Chen, C., Chen, L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z., Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y., Zhang, Y., Ayyampalayam, S., Barkman, T.J., Nguyen, N., Matasci, N., Nelson, D.R., Sayyari, E., Wafula, E.K., Walls, R.L., Warnow, T., An, H., Arrigo, N., Baniaga, A.E., Galuska, S., Jorgensen, S.A., Kidder, T.I., Kong, H., Lu-Irving, P., Marx, H.E., Qi, X., Reardon, C.R., Sutherland, B.L., Tiley, G.P., Welles, S.R., Yu, R., Zhan, S., Gramzow, L., Theißen, G., Wong, G.K.-S., One Thousand Plant Transcriptomes Initiative, 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574, 679–685. https://doi.org/10.1038/s41586-019-1693-2

Lehner, B., Grill, G., 2013. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrological Processes 27, 2171–2186. https://doi.org/10.1002/hyp.9740

Leinonen, R., Sugawara, H., Shumway, M., International Nucleotide Sequence Database Collaboration, 2010. The sequence read archive. Nucleic Acids Research 39, D19–D21.

Lens, F., Jansen, S., Caris, P., Serlet, L., Smets, E., 2005. Comparative wood anatomy of the primuloid clade (Ericales s.l.). Systematic Botany 30, 163–183.

Leroy, T., Louvet, J.-M., Lalanne, C., Le Provost, G., Labadie, K., Aury, J.-M., Delzon, S., Plomion, C., Kremer, A., 2020. Adaptive introgression as a driver of local adaptation to climate in European white oaks. New Phytologist 226, 1171–1182. https://doi.org/10.1111/nph.16095

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Linan, A.G., Lowry II, P.P., Miller, A.J., Schatz, G.E., Sevathian, J.-C., Edwards, C.E., 2020. RAD-sequencing reveals patterns of diversification and hybridization, and the accumulation of reproductive isolation in a clade of partially sympatric, tropical island trees. Molecular Ecology. https://doi.org/10.1111/mec.15736

Linnaeus, C., 1753. Species Plantarum. London.

Liston, A., 2014. 257 nuclear genes for Rosaceae phylogenomics. figshare. Dataset. https://doi.org/10.6084/m9.figshare.1060394.v1

Loiseau, O., Olivares, I., Paris, M., de La Harpe, M., Weigand, A., Koubínová, D., Rolland, J., Bacon, C.D., Balslev, H., Borchsenius, F., Cano, A., Couvreur, T.L.P., Delnatte, C., Fardin, F., Gayot, M., Mejía, F., Mota-Machado, T., Perret, M., Roncal, J., Sanin, M.J., Stauffer, F., Lexer, C., Kessler, M., Salamin, N., 2019. Targeted capture of hundreds of nuclear genes unravels phylogenetic relationships of the diverse Neotropical palm tribe Geonomateae. Frontiers in Plant Science 10, 864. https://doi.org/10.3389/fpls.2019.00864

Lotsy, J.P., 1925. Species or linneon. Genetica 7, 487–506. https://doi.org/10.1007/BF01676287

Löytynoja, A., 2014. Phylogeny-aware alignment with PRANK, in: Multiple Sequence Alignment Methods. Methods in Molecular Biology (Methods and Protocols). Humana Press, Totowa, NJ, 155–170.

Maddison, W.P., 1997. Gene trees in species trees. Systematic Biology 46, 523–536. https://doi.org/10.1093/sysbio/46.3.523

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L.L., Hernández-Hernández, T., 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytologist 207, 437–453.

Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H., Burke, J.M., 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Applications in Plant Sciences 2, apps.1300085. https://doi.org/10.3732/apps.1300085

Manns, U., Anderberg, A.A., 2009. New combinations and names in *Lysimachia* (Myrsinaceae) for species of *Anagallis*, *Pelletiera* and *Trientalis*. Willdenowia 39, 49–54. https://doi.org/10.3372/wi.39.39103

Martin, N.H., Bouck, A.C., Arnold, M.L., 2006. Detecting adaptive trait introgression between *Iris fulva* and *I. brevicaulis* in highly selective field conditions. Genetics 172, 2481. https://doi.org/10.1534/genetics.105.053538

Massana, K.A., Beaulieu, J.M., Matzke, N.J., O'Meara, B.C., 2015. Non-null effects of the null range in biogeographic models: exploring parameter estimation in the DEC model. bioRxiv 026914. https://doi.org/10.1101/026914

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., 2014. Data access for the 1,000 Plants (1KP) project. Gigascience 3, 2047–217X.

Matzke, N.J., 2013. BioGeoBEARS: BioGeography with Bayesian (and likelihood) evolutionary analysis in R Scripts. R package, version 0.2 1, 2013.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20, 1297–1303.

McLay, T.G., Birch, J.L., Gunn, B.F., Ning, W., Tate, J.A., Nauheimer, L., Joyce, E.M., Simpson, L., Schmidt-Lebuhn, A.N., Baker, W.J., 2021. New targets acquired: improving locus recovery from the Angiosperms353 probe set. Applications in Plant Sciences 9.

Memiaghe, H.R., Lutz, J.A., Korte, L., Alonso, A., Kenfack, D., 2016. Ecological importance of small-diameter trees to the structure, diversity and biomass of a tropical evergreen forest at Rabi, Gabon. PLOS One 11, e0154988.

Mez, C., 1902. Myrsinaceae, in: Das Pflanzenreich. Verlag von Wilhelm Engelmann, Leipzig, 1–437.

Milton, T., Assunção, P.A.C.L., Cabello, N., Mori, S., de Oliveira, A.A., Souza, P., Vicentini, A., Dick, C.W., 2022. Biomass and demographic dynamics of the Brazil nut family (Lecythidaceae) in a mature Central Amazon rain forest. Forest Ecology and Management 509, 120058. https://doi.org/10.1016/j.foreco.2022.120058

Minh, B.Q., Nguyen, M.A.T., von Haeseler, A., 2013. Ultrafast approximation for phylogenetic bootstrap. Molecular Biology and Evolution 30, 1188–1195. https://doi.org/10.1093/molbev/mst024

Moquet, L., Vanderplanck, M., Moerman, R., Quinet, M., Roger, N., Michez, D., Jacquemart, A., 2017. Bumblebees depend on ericaceous species to survive in temperate heathlands. Insect Conservation and Diversity 10, 78–93.

Morales-Briones, D.F., Kadereit, G., Tefarikis, D.T., Moore, M.J., Smith, S.A., Brockington, S.F., Timoneda, A., Yim, W.C., Cushman, J.C., Yang, Y., 2020. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae s.l. Systematic Biology. https://doi.org/10.1093/sysbio/syaa066

Mori, S.A., Becker, P., Kincaid, D., 2001. Lecythidaceae of a Central Amazonian lowland forest: implications for conservation, in: Lessons from Amazonia: the ecology and conservation of a fragmented forest. Yale University Press, New Haven, Connecticut, 55–67.

Mori, S.A., Kiernan, E.A., Smith, N.P., Kelly, L.M., Huang, Y.-Y., Prance, G.T., Thiers, B.M., 2017. Observations on the phytogeography of the Lecythidaceae clade (Brazil nut family). Phytoneuron 30, 1–85.

Mori, S.A., Lepsch-Cunha, N., 1995. The Lecythidaceae of a Central Amazonian moist forest. The New York Botanical Garden Press, Bronx, New York.

Mori, S.A., Prance, G.T., 1990. Lecythidaceae, Part 2. The zygomorphic-flowered New World genera (*Couroupita*, *Corythophora*, *Bertholletia*, *Couratari*, *Eschweilera*, & *Lecythis*), Flora Neotropica. New York Botanical Garden Press, Bronx, New York.

Mori, S.A., Smith, N.P., Cornejo, X., Prance, G.T.. 2010 onward. The Lecythidaceae pages. Published on the Internet; http://sweetgum.nybg.org/science/projects/lp/ [Accessed 9 July 2020].

Mori, S.A., Smith, N.P., Huang, Y.-Y., Prance, G.T., Kelly, L.M., Matos, C.C., 2015. Toward a phylogenetic-based generic classification of Neotropical Lecythidaceae—II. Status of *Allantoma*, *Cariniana*, *Couratari*, *Couroupita*, *Grias* and *Gustavia*. Phytotaxa 203, 122–137. https://doi.org/10.11646/phytotaxa.203.2.2

Morton, C.M., Chase, M.W., Kron, K.A., Swensen, S.M., 1996. A molecular evaluation of the monophyly of the order Ebenales based upon rbcL sequence data. Systematic Botany 21, 567–586. https://doi.org/10.2307/2419616

Muñoz, M., Ackerman, J., 2011. Spatial distribution and performance of native and invasive *Ardisia* (Myrsinaceae) species in Puerto Rico: The anatomy of an invasion. Biological Invasions 13, 1543–1558. https://doi.org/10.1007/s10530-010-9912-7

Nazareno, A.G., Dick, C.W., Lohmann, L.G., 2019. A biogeographic barrier test reveals a strong genetic structure for a canopy-emergent amazon tree species. Scientific Reports 9, 18602. https://doi.org/10.1038/s41598-019-55147-1

Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32, 268–274.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., 2001. Terrestrial ecoregions of the world: a new map of life on Earth: a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. BioScience 51, 933–938. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2

Paetzold, C., Wood, K.R., Eaton, D.A.R., Wagner, W.L., Appelhans, M.S., 2019. Phylogeny of Hawaiian *Melicope* (Rutaceae): RAD-seq resolves species relationships and reveals ancient introgression. Frontiers in Plant Science 10, 1074. https://doi.org/10.3389/fpls.2019.01074

Parnell, J., Pedersen, H., Hodkinson, T., Balslev, H., Welzen, P.C., Simpson, D., Middleton, D., Esser, H.-J., Pooma, R., Utteridge, T., Staples, G., 2013. Hybrids and the flora of Thailand. Thai Forest Bulletin 41, 1–9.

Pease, J.B., Brown, J.W., Walker, J.F., Hinchliff, C.E., Smith, S.A., 2018. Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. American Journal of Botany 105, 385–403.

Pease, J.B., Haak, D.C., Hahn, M.W., Moyle, L.C., 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLOS Biology 14, e1002379. https://doi.org/10.1371/journal.pbio.1002379

Pipoly, J.J., Ricketson, J.M., 2000. Discovery of *Ardisia* subgenus *Acrardisia* (Myrsinaceae) in Mesoamerica: another boreotropical element? SIDA, Contributions to Botany 19, 275–283.

Pipoly, J.J., Ricketson, J.M., 1999. Discovery of the Indo-Malesian genus *Hymenandra* (Myrsinaceae) in the Neotropics, and its boreotropical implications. SIDA, Contributions to Botany 18, 701–746.

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M.J., Neale, B., MacArthur, D.G., Banks, E., 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 201178. https://doi.org/10.1101/201178

POWO, 2022. Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet; https://powo.science.kew.org [Accessed 10 January, 2022].

Prance, G.T., Mori, S.A., 1979. Lecythidaceae: Part I: The actinomorphic-flowered New World Lecythidaceae (*Asteranthos*, *Gustavia*, *Grias*, *Allantoma*, & *Cariniana*). Flora Neotropica 21, 1–270.

Prata, E.M.B., Sass, C., Rodrigues, D.P., Domingos, F.M.C.B., Specht, C.D., Damasco, G., Ribas, C.C., Fine, P.V.A., Vicentini, A., 2018. Towards integrative taxonomy in Neotropical botany: disentangling the *Pagamea guianensis* species complex (Rubiaceae). Botanical Journal of the Linnean Society 188, 213–231. https://doi.org/10.1093/botlinnean/boy051

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLOS One 5, e9490.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945.

Pritchard, J.K., Wen, W., Falush, D., 2010. Documentation for STRUCTURE software: version 2. University of Chicago, Chicago, IL.

Puechmaille, S.J., 2016. The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. Molecular Ecology Resources 16, 608–627.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559–575. https://doi.org/10.1086/519795

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., Paterson, A.H., 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biology 20, 1–23.

Quattrocchi, U., 2012. CRC world dictionary of medicinal and poisonous plants: common names, scientific names, eponyms, synonyms, and etymology, 1st ed. CRC Press.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reddy, S., Kimball, R.T., Pandey, A., Hosner, P.A., Braun, M.J., Hackett, S.J., Han, K.-L., Harshman, J., Huddleston, C.J., Kingston, S., 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Systematic Biology 66, 857–879.

Ree, R.H., Smith, S.A., 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Systematic Biology 57, 4–14. https://doi.org/10.1080/10635150701883881

Revell, L.J., 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3, 217–223.

Richards, E.J., Brown, J.M., Barley, A.J., Chong, R.A., Thomson, R.C., 2018. Variation across mitochondrial gene trees provides evidence for systematic error: how much gene tree variation is biological? Systematic biology 67, 847–860.

Rieseberg, L.H., Soltis, D., 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evolutionary Trends in Plants 5, 65–84.

Rieseberg, L.H., Wendel, J.F., 1993. Introgression and its consequences in plants, in: Hybrid zones and the evolutionary process. Oxford University Press, New York, NY.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53, 131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Rose, J.P., Kleist, T.J., Löfstrand, S.D., Drew, B.T., Schönenberger, J., Sytsma, K.J., 2018. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. Molecular Phylogenetics and Evolution 122, 59–79. https://doi.org/10.1016/j.ympev.2018.01.014

Rothfels, C.J., Johnson, A.K., Hovenkamp, P.H., Swofford, D.L., Roskam, H.C., Fraser-Jenkins, C.R., Windham, M.D., Pryer, K.M., 2015. Natural hybridization between genera that diverged from each other approximately 60 million years ago. The American Naturalist 185, 433–442.

Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution 19, 101–109. https://doi.org/10.1093/oxfordjournals.molbev.a003974

Schley, R.J., Pennington, R.T., Pérez-Escobar, O.A., Helmstetter, A.J., de la Estrella, M., Larridon, I., Sabino Kikuchi, I.A.B., Barraclough, T.G., Forest, F., Klitgård, B., 2020. Introgression across evolutionary scales suggests reticulation contributes to Amazonian tree diversity. Molecular Ecology 29, 4170–4185. https://doi.org/10.1111/mec.15616

Schmid, M., 2011. Contribution to the knowledge of Primulaceae (ex Myrsinaceae) of New Caledonia. III. *Tapeinosperma* Hook. f. and *Mangenotiella* gen. nov. genera. Adansonia 34, 279–341.

Schmitt, S., Tysklind, N., Derroire, G., Heuertz, M., Hérault, B., 2021. Topography shapes the local coexistence of tree species within species complexes of Neotropical forests. Oecologia 196, 389–398. https://doi.org/10.1007/s00442-021-04939-2

Schönenberger, J., Anderberg, A.A., Sytsma, K.J., 2005. Molecular phylogenetics and patterns of floral evolution in the Ericales. International Journal of Plant Sciences 166, 265–288. https://doi.org/10.1086/427198

Seo, T.-K., 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Molecular Biology and Evolution 25, 960–971. https://doi.org/10.1093/molbev/msn043

Shen, X.-X., Hittinger, C.T., Rokas, A., 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecology & Evolution 1, 1–10.

Shi, T., Huang, H., Barker, M.S., 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. Annals of Botany 106, 497–504.

Sievert, C., 2020. Interactive web-based data visualization with R, plotly, and shiny, Chapman & Hall/CRC The R Series. CRC Press, Taylor and Francis Group.

Simon, S., Narechania, A., DeSalle, R., Hadrys, H., 2012. Insect phylogenomics: exploring the source of incongruence using new transcriptomic data. Genome Biology and Evolution 4, 1295–1309.

Slater, G.S.C., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31. https://doi.org/10.1186/1471-2105-6-31

Smith, S.A., Donoghue, M.J., 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322, 86–89. https://doi.org/10.1126/science.1163197

Smith, S.A., Moore, M.J., Brown, J.W., Yang, Y., 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evolutionary Biology 15, 1–15.

Smith, S.A., O'Meara, B.C., 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics 28, 2689–2690. https://doi.org/10.1093/bioinformatics/bts492

Smith, S.A., Walker-Hale, N., Walker, J.F., Brown, J.W., 2020. Phylogenetic conflicts, combinability, and deep phylogenomics in plants. Systematic Biology 69, 579–592.

Solís-Lemus, C., Bastide, P., Ané, C., 2017. PhyloNetworks: a package for phylogenetic networks. Molecular Biology and Evolution 34, 3292–3298. https://doi.org/10.1093/molbev/msx235

Soza, V.L., Lindsley, D., Waalkes, A., Ramage, E., Patwardhan, R.P., Burton, J.N., Adey, A., Kumar, A., Qiu, R., Shendure, J., 2019. The *Rhododendron* genome and chromosomal organization provide insight into shared whole-genome duplications across the heath family (Ericaceae). Genome Biology and Evolution 11, 3353–3371.

Springer, M.S., Gatesy, J., 2016. The gene tree delusion. Molecular Phylogenetics and Evolution 94, 1–33. https://doi.org/10.1016/j.ympev.2015.07.018

Ståhl, B., 2010. Theophrastaceae. Flora Neotropica 105, 1–161.

Ståhl, B., 2004. Samolaceae, in: Kubitzki, K. (Ed.), Flowering Plants · Dicotyledons: Celastrales, Oxalidales, Rosales, Cornales, Ericales. Springer Berlin Heidelberg, Berlin, Heidelberg, 387–389.

Ståhl, B., Anderberg, A.A., 2004. Myrsinaceae, in: Kubitzki, K. (Ed.), Flowering Plants · Dicotyledons: Celastrales, Oxalidales, Rosales, Cornales, Ericales. Springer Berlin Heidelberg, Berlin, Heidelberg, 266–281. https://doi.org/10.1007/978-3-662-07257-8_30

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stevens, P.F. 2001 onward. Angiosperm phylogeny website, version 14, July 2017 [more or less continuously updated]. Published on the Internet; http://www.mobot.org/MOBOT/research/APweb/ [Accessed 11 July 2019].

Stretton, A.O.W., 2002. The first sequence. Fred Sanger and insulin. Genetics 162, 527–532. https://doi.org/10.1093/genetics/162.2.527

Strijk, J.S., Bone, R.E., Thébaud, C., Buerki, S., Fritsch, P.W., Hodkinson, T.R., Strasberg, D., 2014. Timing and tempo of evolutionary diversification in a biodiversity hotspot: Primulaceae on Indian Ocean islands. Journal of Biogeography 41, 810–822.

Stull, G.W., Soltis, P.S., Soltis, D.E., Gitzendanner, M.A., Smith, S.A., 2020. Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. American Journal of Botany 107, 790–805.

Suarez-Gonzalez, A., Lexer, C., Cronk, Q.C.B., 2018. Adaptive introgression: a plant perspective. Biology Letters 14, 20170688. https://doi.org/10.1098/rsbl.2017.0688

Sumanon, P., Eiserhardt, W.L., Balslev, H., Utteridge, T.M., 2021. Six new species of *Maesa* (Primulaceae) from Papua New Guinea. Phytotaxa 505, 245–261.

Tange, O., 2011. Gnu parallel-the command-line power tool. The USENIX Magazine 36, 42–47. https://doi.org/10.5281/zenodo.16303

Ter Steege, H., Pitman, N.C.A., Phillips, O.L., Chave, J., Sabatier, D., Duque, A., Molino, J.-F., Prévost, M.-F., Spichiger, R., Castellanos, H., von Hildebrand, P., Vásquez, R., 2006. Continental-scale patterns of canopy tree composition and function across Amazonia. Nature 443, 444–447. https://doi.org/10.1038/nature05134

Ter Steege, H., Pitman, N.C.A., Sabatier, D., Baraloto, C., Salomão, R.P., Guevara, J.E., Phillips, O.L., Castilho, C.V., Magnusson, W.E., Molino, J.-F., Monteagudo, A., Núñez Vargas, P., Montero, J.C., Feldpausch, T.R., Coronado, E.N.H., Killeen, T.J., Mostacedo, B., Vasquez, R., Assis, R.L., Terborgh, J., Wittmann, F., Andrade, A., Laurance, W.F., Laurance, S.G.W., Marimon, B.S., Marimon, B.-H., Guimarães Vieira, I.C., Amaral, I.L., Brienen, R., Castellanos, H., Cárdenas López, D., Duivenvoorden, J.F., Mogollón, H.F., Matos, F.D. de A., Dávila, N., García-Villacorta, R., Stevenson Diaz, P.R., Costa, F., Emilio, T., Levis, C., Schietti, J., Souza, P., Alonso, A., Dallmeier, F., Montoya, A.J.D., Fernandez Piedade, M.T., Araujo-Murakami, A., Arroyo, L., Gribel, R., Fine, P.V.A., Peres, C.A., Toledo, M., Aymard C., G.A., Baker, T.R., Cerón, C., Engel, J., Henkel, T.W., Maas, P., Petronelli, P., Stropp, J., Zartman, C.E., Daly, D., Neill, D., Silveira, M., Paredes, M.R., Chave, J., Lima Filho, D. de A., Jørgensen, P.M., Fuentes, A., Schöngart, J., Cornejo Valverde, F., Di Fiore, A., Jimenez, E.M., Peñuela Mora, M.C., Phillips, J.F., Rivas, G., van Andel, T.R., von Hildebrand, P., Hoffman, B., Zent, E.L., Malhi, Y., Prieto, A., Rudas, A., Ruschell, A.R., Silva, N., Vos, V., Zent, S., Oliveira, A.A., Schutz, A.C., Gonzales, T., Trindade Nascimento, M., Ramirez-Angulo, H., Sierra, R., Tirado, M., Umaña Medina, M.N., van der Heijden, G., Vela, C.I.A., Vilanova Torre, E., Vriesendorp, C., Wang, O., Young, K.R., Baider, C., Balslev, H., Ferreira, C., Mesones, I., Torres-Lezama, A., Urrego Giraldo, L.E., Zagt, R., Alexiades, M.N., Hernandez, L., Huamantupa-Chuquimaco, I., Milliken, W., Palacios Cuenca, W., Pauletto, D., Valderrama Sandoval, E., Valenzuela Gamarra, L., Dexter, K.G., Feeley, K., Lopez-Gonzalez, G., Silman, M.R., 2013. Hyperdominance in the Amazonian tree flora. Science 342, 1243092. https://doi.org/10.1126/science.1243092

Ter Steege, H., Prado, P.I., Lima, R.A.F. de, Pos, E., de Souza Coelho, L., de Andrade Lima Filho, D., Salomão, R.P., Amaral, I.L., de Almeida Matos, F.D., Castilho, C.V., Phillips, O.L., Guevara, J.E., de Jesus Veiga Carim, M., Cárdenas López, D., Magnusson, W.E., Wittmann, F., Martins, M.P., Sabatier, D., Irume, M.V., da Silva Guimarães, J.R., Molino, J.-F., Bánki, O.S., Piedade, M.T.F., Pitman, N.C.A., Ramos, J.F., Monteagudo Mendoza, A., Venticinque, E.M., Luize, B.G., Núñez Vargas, P., Silva, T.S.F., de Leão Novo, E.M.M., Reis, N.F.C., Terborgh, J., Manzatto, A.G., Casula, K.R., Honorio Coronado, E.N., Montero, J.C., Duque, A., Costa, F.R.C., Castaño Arboleda, N., Schöngart, J., Zartman, C.E., Killeen, T.J., Marimon, B.S., Marimon-Junior, B.H., Vasquez, R., Mostacedo, B., Demarchi, L.O., Feldpausch, T.R., Engel, J., Petronelli, P., Baraloto, C., Assis, R.L., Castellanos, H., Simon, M.F., de Medeiros, M.B., Quaresma, A., Laurance, S.G.W., Rincón, L.M., Andrade, A., Sousa, T.R., Camargo, J.L., Schietti, J., Laurance, W.F., de Queiroz, H.L., Nascimento, H.E.M., Lopes, M.A., de Sousa Farias, E., Magalhães, J.L.L., Brienen, R., Aymard C., G.A., Revilla, J.D.C., Vieira, I.C.G., Cintra, B.B.L., Stevenson, P.R., Feitosa, Y.O., Duivenvoorden, J.F., Mogollón, H.F., Araujo-Murakami, A., Ferreira, L.V., Lozada, J.R., Comiskey, J.A., de Toledo, J.J., Damasco, G., Dávila, N., Lopes, A., García-Villacorta, R., Draper, F., Vicentini, A., Cornejo Valverde, F., Lloyd, J., Gomes, V.H.F., Neill, D., Alonso, A., Dallmeier, F., de Souza, F.C., Gribel, R., Arroyo, L., Carvalho, F.A., de Aguiar, D.P.P., do Amaral, D.D., Pansonato, M.P., Feeley, K.J., Berenguer, E., Fine, P.V.A., Guedes, M.C., Barlow, J., Ferreira, J., Villa, B., Peñuela Mora, M.C., Jimenez, E.M., Licona, J.C., Cerón, C., Thomas, R., Maas, P., Silveira, M., Henkel, T.W., Stropp, J., Paredes, M.R., Dexter, K.G., Daly, D., Baker, T.R., Huamantupa-Chuquimaco, I., Milliken, W., Pennington, T., Tello, J.S., Pena, J.L.M., Peres, C.A., Klitgaard, B., Fuentes, A., Silman, M.R., Di Fiore, A., von Hildebrand, P., Chave, J., van Andel, T.R., Hilário, R.R., Phillips, J.F., Rivas-Torres, G., Noronha, J.C., Prieto, A., Gonzales, T., de Sá Carpanedo, R., Gonzales, G.P.G., Gómez, R.Z., de Jesus Rodrigues, D., Zent, E.L., Ruschel, A.R., Vos, V.A., Fonty, É., Junqueira, A.B., Doza, H.P.D., Hoffman, B., Zent, S., Barbosa, E.M., Malhi, Y., de Matos Bonates, L.C., de Andrade Miranda, I.P., Silva, N., Barbosa, F.R., Vela, C.I.A., Pinto, L.F.M., Rudas, A., Albuquerque, B.W., Umaña, M.N., Carrero Márquez, Y.A., van der Heijden, G., Young, K.R., Tirado, M., Correa, D.F., Sierra, R., Costa, J.B.P., Rocha, M., Vilanova Torre, E., Wang, O., Oliveira, A.A., Kalamandeen, M., Vriesendorp, C., Ramirez-Angulo, H., Holmgren, M., Nascimento, M.T., Galbraith, D., Flores, B.M., Scudeller, V.V., Cano, A., Ahuite Reategui, M.A., Mesones, I., Baider, C., Mendoza, C., Zagt, R., Urrego Giraldo, L.E., Ferreira, C., Villarroel, D., Linares-Palomino, R., Farfan-Rios, W., Farfan-Rios, W., Casas, L.F., Cárdenas, S., Balslev, H., Torres-Lezama, A., Alexiades, M.N., Garcia-Cabrera, K., Valenzuela Gamarra, L., Valderrama Sandoval, E.H., Ramirez Arevalo, F., Hernandez, L., Sampaio, A.F., Pansini, S., Palacios Cuenca, W., de Oliveira, E.A., Pauletto, D., Levesley, A., Melgaço, K., Pickavance, G., 2020. Biased-corrected richness estimates for the Amazonian tree flora. Scientific Reports 10, 10130. https://doi.org/10.1038/s41598-020-66686-3

Than, C., Ruths, D., Nakhleh, L., 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9, 322. https://doi.org/10.1186/1471-2105-9-322

Utteridge, T.M.A., 2012. Four new species of *Maesa* Forssk. (Primulaceae) from Malesia. Kew Bulletin 67, 367–378. https://doi.org/10.1007/s12225-012-9383-3

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics 43, 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Vargas, O.M., Dick, C.W., 2020. Diversification history of Neotropical Lecythidaceae, an ecologically dominant tree family of Amazon Rain Forest, in: Rull, V., Carnaval, A.C. (Eds.), Neotropical Diversification: Patterns and Processes. Springer International Publishing, Cham, 791–809. https://doi.org/10.1007/978-3-030-31167-4_29

Vargas, O.M., Goldston, B., Grossenbacher, D.L., Kay, K.M., 2020. Patterns of speciation are similar across mountainous and lowland regions for a Neotropical plant radiation (Costaceae: *Costus*). Evolution 74, 2644–2661. https://doi.org/10.1111/evo.14108

Vargas, O.M., Heuertz, M., Smith, S.A., Dick, C.W., 2019. Target sequence capture in the Brazil nut family (Lecythidaceae): marker selection and in silico capture from genome skimming data. Molecular Phylogenetics and Evolution 135, 98–104. https://doi.org/10.1016/j.ympev.2019.02.020

Vargas, O.M., Ortiz, E.M., Simpson, B.B., 2017. Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). New Phytologist 214, 1736–1750.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., 2001. The sequence of the human genome. Science 291, 1304–1351.

Walker, J.D., Geissman, J., Bowring, S., Babcock, L., 2013. The Geological Society of America geologic time scale. GSA Bulletin 125, 259–272.

Walker, J.F., Brown, J.W., Smith, S.A., 2018a. Analyzing contentious relationships and outlier genes in phylogenomics. Systematic Biology 67, 916–924. https://doi.org/10.1093/sysbio/syy043

Walker, J.F., Walker-Hale, N., Vargas, O.M., Larson, D.A., Stull, G.W., 2019. Characterizing gene tree conflict in plastome-inferred phylogenies. PeerJ 7, e7747. https://doi.org/10.7717/peerj.7747

Walker, J.F., Yang, Y., Feng, T., Timoneda, A., Mikenas, J., Hutchison, V., Edwards, C., Wang, N., Ahluwalia, S., Olivieri, J., 2018b. From cacti to carnivores: improved phylotranscriptomic sampling and hierarchical homology inference provide further insight into the evolution of Caryophyllales. American Journal of Botany 105, 446–462.

Walker, J.F., Yang, Y., Moore, M.J., Mikenas, J., Timoneda, A., Brockington, S.F., Smith, S.A., 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. American Journal of Botany 104, 858–867.

Wang, N., Yang, Y., Moore, M.J., Brockington, S.F., Walker, J.F., Brown, J.W., Liang, B., Feng, T., Edwards, C., Mikenas, J., 2019. Evolution of Portulacineae marked by gene tree conflict and gene family expansion associated with adaptation to harsh environments. Molecular Biology and Evolution 36, 112–126.

Wanntorp, L., Ronse De Craene, L., Peng, C.-I., Anderberg, A.A., 2012. Floral ontogeny and morphology of *Stimpsonia* and *Ardisiandra*, two aberrant genera of the primuloid clade of Ericales. International Journal of Plant Sciences 173, 1023–1035. https://doi.org/10.1086/667607

Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., Xia, E., Lu, Y., Tai, Y., She, G., 2018. Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. Proceedings of the National Academy of Sciences 115, E4151–E4158.

Whitney, K.D., Randell, R.A., Rieseberg, L.H., 2010. Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. New Phytologist 187, 230–239. https://doi.org/10.1111/j.1469-8137.2010.03234.x

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences 111, E4859–E4868.

Wickham, H., 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York.

Wu, H., Ma, T., Kang, M., Ai, F., Zhang, J., Dong, G., Liu, J., 2019. A high-quality *Actinidia chinensis* (kiwifruit) genome. Horticulture Research 6, 117. https://doi.org/10.1038/s41438-019-0202-y

Xia, E.-H., Zhang, H.-B., Sheng, J., Li, K., Zhang, Q.-J., Kim, C., Zhang, Y., Liu, Y., Zhu, T., Li, W., Huang, H., Tong, Y., Nan, H., Shi, Cong, Shi, Chao, Jiang, J.-J., Mao, S.-Y., Jiao, J.-Y., Zhang, D., Zhao, Y., Zhao, Y.-J., Zhang, L.-P., Liu, Y.-L., Liu, B.-Y., Yu, Y., Shao, S.-F., Ni, D.-J., Eichler, E.E., Gao, L.-Z., 2017. The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Molecular Plant 10, 866–877. https://doi.org/10.1016/j.molp.2017.04.002

Yang, C.-J., Hu, J.-M., 2022. Molecular phylogeny of Asian *Ardisia* (Myrsinoideae, Primulaceae) and their leaf-nodulated endosymbionts, *Burkholderia* s.l. (Burkholderiaceae). PLOS ONE 17, e0261188. https://doi.org/10.1371/journal.pone.0261188

Yang, Y., Moore, M.J., Brockington, S.F., Mikenas, J., Olivieri, J., Walker, J.F., Smith, S.A., 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. New Phytologist 217, 855–870.

Yang, Y., Moore, M.J., Brockington, S.F., Soltis, D.E., Wong, G.K.-S., Carpenter, E.J., Zhang, Y., Chen, L., Yan, Z., Xie, Y., 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Molecular Biology and Evolution 32, 2001–2014.

Yang, Y., Smith, S.A., 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Molecular Biology and Evolution 31, 3081–3092. https://doi.org/10.1093/molbev/msu245

Yang, Z., 2014. Molecular evolution: a statistical approach. Oxford University Press.

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 153. https://doi.org/10.1186/s12859-018-2129-y

Zhang, C., Zhang, T., Luebert, F., Xiang, Y., Huang, C.-H., Hu, Y., Rees, M., Frohlich, M.W., Qi, J., Weigend, M., Ma, H., 2020. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. Molecular Biology and Evolution 37, 3188–3210. https://doi.org/10.1093/molbev/msaa160

Zhang, Q., Zhao, L., Folk, R.A., Zhao, J.-L., Zamora, N.A., Yang, S.-X., Soltis, D.E., Soltis, P.S., Gao, L.-M., Peng, H., 2022. Phylotranscriptomics of Theaceae: generic level relationships, reticulation and whole-genome duplication. Annals of botany.

Zhbannikov, I.Y., Hunter, S.S., Foster, J.A., Settles, M.L., 2017. SeqyClean: a pipeline for high-throughput sequence data preprocessing. Presented at the Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 407–416.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., Weir, B.S., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Zhou, L., Lin, Y., Feng, B., Zhao, J., Tang, J., 2017. Phylogeny analysis from gene-order data with massive duplications. BMC Genomics 18, 760. https://doi.org/10.1186/s12864-017-4129-0

Zuckerkandl, E., Pauling, L., 1965. Molecules as documents of evolutionary history. Journal of Theoretical Biology 8, 357–366. https://doi.org/10.1016/0022-5193(65)90083-4

Zwaenepoel, A., Van de Peer, Y., 2019. Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. Molecular Biology and Evolution 36, 1384–1404.