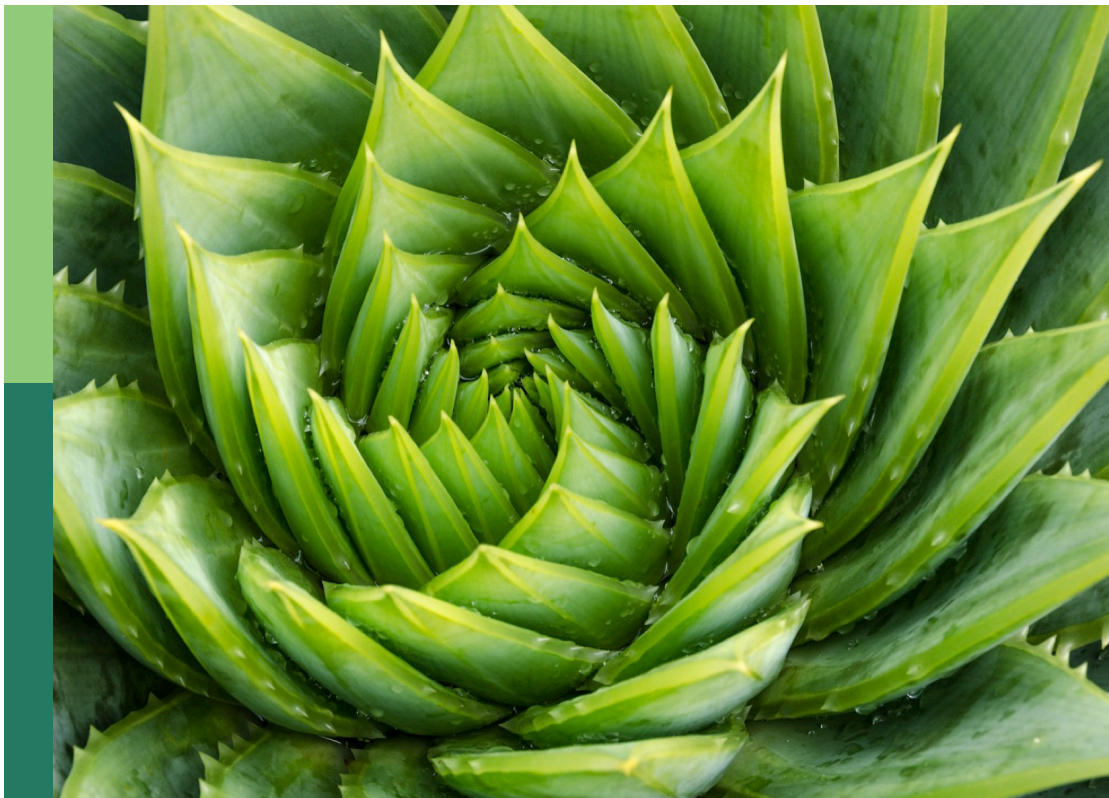
frontiers | Research Topics

Phylogenomic discordance in plant systematics

Edited by
Stefan Wanke and Susann Wicke

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-3899-9
DOI 10.3389/978-2-8325-3899-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Phylogenomic discordance in plant systematics

Topic editors

Stefan Wanke — Technical University Dresden, Germany

Susann Wicke — Humboldt University of Berlin, Germany

Citation

Wanke, S., Wicke, S., eds. (2023). *Phylogenomic discordance in plant systematics*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3899-9

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 27 October 2023
DOI 10.3389/fpls.2023.1308126

OPEN ACCESS

EDITED AND REVIEWED BY
Gerald Matthias Schneeweiss,
University of Vienna, Austria

*CORRESPONDENCE
Susann Wicke
susann.wicke@hu-berlin.de

RECEIVED 05 October 2023
ACCEPTED 20 October 2023
PUBLISHED 27 October 2023

CITATION
Wanke S and Wicke S (2023) Editorial: Phylogenomic discordance in plant systematics.
Front. Plant Sci. 14:1308126.
doi: 10.3389/fpls.2023.1308126

Editorial: Phylogenomic discordance in plant systematics

Stefan Wanke [1,2] and Susann Wicke [3]*

[1]Institute for Botany, Technische Universität Dresden, Dresden, Germany, [2]Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3]Institute for Biology, Humboldt-Universität zu Berlin, Berlin, Germany

KEYWORDS

gene tree, species tree, incomplete lineage sorting, plastid (chloroplast) DNA, phylogenetic discordance, phylogenomic analysis, DNA barcoding, reticulate evolution

Editorial on the Research Topic
Phylogenomic discordance in plant systematics

In the *omics*-age of molecular systematics, entire genomes or genomic segments often contradict each other or the broader consensus of organismal relationships. Furthermore, this potentially conflicts with pre-*omics* phylogenetics, where true conflicts often remained undiscovered due to data limitations. Unlike earlier times when adding markers and taxa might have helped to address or even (superficially) resolve conflicts around phylogenetic hypotheses, incongruences arising from genome inferences remain challenging. Phylogenetic discordance can result from various evolutionary processes that lead to disparities between gene trees and species trees. This extends to genomic discordance, where organellar and nuclear genomes exhibit different coalescent paths or phenomena such as organelle capture. Advances in analytical methods enable the comparison of hundreds to thousands of loci across all plant genomes, offering a comprehensive view of phylogenomic complexity. This Research Topic explores high-throughput sequence-based phylogenomic studies that uncover discordant phylogenies. A total of 79 authors present a rich array of 14 original research articles focusing on phylogenomic studies based on high-throughput sequence data. These studies delve into the discordant phylogenies between, among, or within organellar genomes and the nuclear genome.

Through large-scale comparative analysis of over 3,600 plant plastomes, Yang et al. contribute to the growing body of studies that find potential issues with phylogenetic inference using plastid-only data, which suffer from saturation at third codon positions. Similarly, organellar phylogenomic analysis by Wu et al. of Poales, one of the largest monocot orders, attribute phylogenetic conflicts to potential ancient rapid radiation, advocating the integration of nuclear data to fully resolve relationships. In fact, a set of articles caution against overreliance on organellar genomes in resolving evolutionary relationships, due to their intrinsic limitations. Low mutation rates, extensive homoplasy, and lack of taxonomic coherence limit the utility of plastid genomes, for example in *Salix* spp. (Salicaceae), as Wagner et al. demonstrate through their analysis of shrub willow plastomes and comparing these to RAD sequencing-based data. They suggest nuclear data may better resolve biogeographical questions, as these reflect not just one coalescence line.

Nuclear genomes have their own challenges. Wu et al. found substantial phylogenetic conflicts within the plastid genomes of the Poales, as well as among the plastid, mitochondrial, and nuclear data, suggesting a complicated evolutionary history with rapid radiation and polyploidy, e.g. through hybridization. Such findings lend credence to calls by Jost et al. and Kandziora et al. for integrating evidence across genomic compartments. While organellar genomes have proven value in DNA-barcoding applications, resolving deep phylogenies may require more judicious data integration. Garrett et al. point out extensive gene tree conflicts in *Euphrasia* spp. (Orobanchaceae), limiting the utility of genome skimming for species identification. Such factors underscore the need for robust practices as phylogenomic datasets grow in scale. Complementing this topical complex, Hernández-Gutiérrez et al. advocate considering rate heterogeneity across loci and applying sorting approaches to mitigate its confounding effects. Thureborn et al. used the normalized quartet score (NQS) to assess gene tree discordance for the coffee family Rubiaceae, and Hodel et al. employed network analysis to examine phylogenetic discordance in their study of the apple tribe (Maleae, Rosaceae). The insights from these methodological approaches inform efforts to analyze phylogenomic datasets. Gene tree estimation and gene tree/species tree reconciliation practices also warrant scrutiny, with Kandziora et al. noting the potential of paralogy to mislead phylogenetic inference.

Resolving phylogenetic relationships within plant lineages where rapid diversification occurred millions of years ago can be particularly challenging for several reasons, incl. limited genetic variation, incomplete lineage sorting (ILS), hybridization, long branch attraction, limited fossil record, lack of informative characters, or complex evolutionary processes such as adaptive radiation, where species rapidly adapt to exploit different ecological niches. These processes can result in intricate patterns of diversification that are challenging to unravel. Among such phylogenetically challenging groups of plants are sages (*Salvia* spp., Lamiaceae), comprising approximately a thousand species. Rose et al. and Lara-Cabrera et al. both used a combination of nuclear and plastid data obtained from hybrid enrichment and off-target plastome sequences to infer gene and species phylogenies by Bayesian and Maximum Likelihood (ML) multispecies coalescent-based approaches. To examine the concordance and discordance among nuclear loci and between the nuclear and plastid genomes in detail, simulations were run to test whether ILS underlies the phylogenetic discordance (Rose et al.) and to infer the robustness of inferences in light of varying extents of missing data (Lara-Cabrera et al.). Together, these studies provide a well-supported backbone species tree of *Salvia* spp. across phylogenetic scales and genomes, suggesting that past difficulties in inferring relationships may have been caused by a combination of uninformative markers, ILS, and horizontal gene flow.

ILS arising from rapid radiations is also identified as major potential driver of phylogenomic discordance by Zheng et al. in their work on the *Quercus franchetii* complex (Fagaceae) spanning the Himalaya region since the Oligocene. They suggest that tectonic shifts and environmental heterogeneity have promoted allopatric speciation, restricting gene flow. This could have increased the chance of ILS, although the hypothesis of an ancient rapid diversification in the group remains to be tested. Likewise, Hernández-Gutiérrez et al. find short branches and incongruent relationships between Malvaceae lineages, indicating potential ILS during diversification. Using triplet analysis the study found that the signal of ILS can be obscured by even low levels of introgression. This underscores the need for robust methods like gene tree sorting and topology weighting, applied by Jost et al. and Kandziora et al. in Piperales and *Loricaria* (Asteraceae), respectively. Such approaches can provide greater confidence in elucidating whether ILS alone or complex factors underlie phylogenetic discordance.

A predominant theme emerging is the potential role of reticulate evolutionary processes like hybridization and introgression as contributors to phylogenomic discordance. Hernández-Gutiérrez et al. present evidence of introgression contributing to discordance on top of ILS-related conflicts between subfamilies in Malvaceae. In contrast, Liu et al. in their study on the *Pedicularis siphonantha* complex (Orobanchaceae), endemic to Southwest China, implicate ancient hybridization events in shaping the topological conflicts observed between nuclear and plastid phylogenies. Similarly, Hsieh et al., through their analysis of 93 plastid genomes representing all genera of Berberidaceae, suggest that ancient hybridization between diverging lineages gave rise to intermediate genera like *Alloberberis*. They note substantial sequence variation in plastid markers among species, thereby highlighting plastomic fluidity. While these specific cases lend evidence for hybridization's influence, its pervasiveness and evolutionary importance across diverse plant families require further investigation through rigorous assessments to avoid overstating its role.

In terms of implications, Hsieh et al., Rose et al., and Garrett et al. note that extensive phylogenetic discordance poses challenges for taxonomy, species delimitation, and DNA barcoding efforts in diverse plant groups. Extended barcodes may have limited utility in taxa exhibiting high gene tree conflicts. Hernández-Gutiérrez et al. posit that resolving deep phylogenetic uncertainties may require moving beyond just amassing larger genomic datasets, to focusing on data quality and model adequacy. Meanwhile, Yang et al. suggest dense taxon sampling may not always improve phylogenetic accuracy in the face of pervasive ILS. Such perspectives serve as important reminders that more data does not automatically equate to simpler evolutionary interpretations. It also (re)opens an exciting debate on how plant classification can develop under coexisting phylogenetic hypotheses that potentially arise from (currently) unresolvable topological conflicts among large sets of gene trees.

In synthesizing these findings, a central theme emerges: the widespread occurrence of phylogenetic discordance, arising from a complex interplay of biological and methodological factors. Factors such as reticulate evolution, incomplete lineage sorting, and rapid radiations all contribute to the intricate tapestry of evolutionary histories. To move forward, we must prioritize the development of robust comparative methods and study designs that harness the power of genomic data. It is crucial to approach phylogenetic conflicts with care, integrating evidence from various data types while acknowledging

the heterogeneity among (sub-)genomic regions. Embracing the intricacies unveiled through phylogenomics grants us deeper insights into the mechanisms driving plant diversity. However, this expanding body of knowledge should also foster humility as we increasingly appreciate the multifaceted nature of evolutionary narratives.

Author contributions

SWa: Conceptualization, Writing – original draft. SWi: Conceptualization, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

frontiers in Plant Science

ORIGINAL RESEARCH
published: 12 April 2021
doi: 10.3389/fpls.2021.642598

Discordant Phylogenomic Placement of Hydnoraceae and Lactoridaceae Within Piperales Using Data From All Three Genomes

Matthias Jost[1], Marie-Stéphanie Samain[2], Isabel Marques[3,4], Sean W. Graham[3] and Stefan Wanke[1]**

[1] Institut für Botanik, Technische Universität Dresden, Dresden, Germany, [2] Instituto de Ecología, A.C., Red de Diversidad Biológica del Occidente Mexicano, Pátzcuaro, Mexico, [3] Department of Botany, University of British Columbia, Vancouver, BC, Canada, [4] Plant-Environment Interactions and Biodiversity Lab, Forest Research Centre, Instituto Superior de Agronomia, Universidade de Lisboa, Lisbon, Portugal

OPEN ACCESS

Edited by:
Hervé Sauquet,
Royal Botanic Gardens and Domain Trust, Australia

Reviewed by:
Daniel Lee Nickrent,
Southern Illinois University Carbondale, United States
Gregory W. Stull,
Kunming Institute of Botany, Chinese Academy of Sciences, China

***Correspondence:**
Matthias Jost
matthias.jost@tu-dresden.de
Stefan Wanke
stefan.wanke@tu-dresden.de

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: 16 December 2020
Accepted: 17 March 2021
Published: 12 April 2021

Citation:
Jost M, Samain M-S, Marques I, Graham SW and Wanke S (2021) Discordant Phylogenomic Placement of Hydnoraceae and Lactoridaceae Within Piperales Using Data From All Three Genomes. Front. Plant Sci. 12:642598. doi: 10.3389/fpls.2021.642598

Phylogenetic relationships within the magnoliid order Piperales have been studied extensively, yet the relationships of the monotypic family Lactoridaceae and the holoparasitic Hydnoraceae to the remainder of the order remain a matter of debate. Since the first confident molecular phylogenetic placement of Hydnoraceae among Piperales, different studies have recovered various contradictory topologies. Most phylogenetic hypotheses were inferred using only a few loci and have had incomplete taxon sampling at the genus level. Based on these results and an online survey of taxonomic opinion, the Angiosperm Phylogeny Group lumped both Hydnoraceae and Lactoridaceae in Aristolochiaceae; however, the latter family continues to have unclear relationships to the aforementioned taxa. Here we present extensive phylogenomic tree reconstructions based on up to 137 loci from all three subcellular genomes for all genera of Piperales. We infer relationships based on a variety of phylogenetic methods, explore instances of phylogenomic discordance between the subcellular genomes, and test alternative topologies. Consistent with these phylogenomic results and a consideration of the principles of phylogenetic classification, we propose to exclude Hydnoraceae and Lactoridaceae from the broad circumscription of Aristolochiaceae, and instead favor recognition of four monophyletic and morphologically well circumscribed families in the perianth-bearing Piperales: Aristolochiaceae, Asaraceae, Hydnoraceae, and Lactoridaceae, with a total of six families in the order.

Keywords: Aristolochiaceae, *Hydnora*, *Prosopanche*, *Lactoris*, *Verhuellia*, plastome, mitochondrial, nuclear

INTRODUCTION

The magnoliid clade Piperales represents the largest angiosperm order outside the eudicots and monocots, as it includes some 4,200 species in 16 genera (Meng et al., 2002; Quijano-Abril et al., 2006; Wanke et al., 2006; Oelschlägel et al., 2011; Wagner et al., 2012; Frenzke et al., 2015; Sinn et al., 2015; Bolin et al., 2018; Funez et al., 2019). Members of this major angiosperm lineage,

with an estimated crown diversification of (174-)148(-124) Myr (Salomo et al., 2017) have a nearly worldwide distribution and are present in most terrestrial biomes, occurring from sea level to high mountain areas above the tree line. The order is the most morphologically diverse magnoliid lineage (Isnard et al., 2012), comprising nearly all growth and life forms, including geophytes, epiphytes, annuals, perennials, herbs, succulents, shrubs, trees, lianas, aquatic plants, and parasites (Wanke et al., 2007a; Isnard et al., 2012). In addition, their floral morphology is extremely diverse, ranging from reduced perianth-less, and likely wind-pollinated flowers in Piperaceae and Saururaceae, to insect-trapping flowers in, for example, Aristolochiaceae, and extremely modified (and at least partially subterranean) beetle-pollinated flowers in Hydnoraceae (Bolin et al., 2009; Oelschlägel et al., 2009, 2015; Seymour et al., 2009). Piperales have been the subject of extensive studies in a broad range of scientific fields, including pharmacological investigations of *Aristolochia* (Sati et al., 2011), *Piper* (Zaveri et al., 2010; Ahmad et al., 2012; Shah et al., 2016), *Peperomia* (Hamid et al., 2007), and *Thottea* (Raju and Ramesh, 2012), and investigations into the repelling properties of essential oils of certain *Piper* species to fire ants (Souto et al., 2012), the cattle tick (Silva et al., 2009), and other arthropods (Mamood et al., 2017). Other studies have focused on their conservation biology (Stuessy et al., 1992; Ricci, 2001), pollination biology (Oelschlägel et al., 2016), floral development (Jaramillo et al., 2004; Samain et al., 2010; Pabón-Mora et al., 2015, 2020), the evolution of epiphytism and fruit traits (Frenzke et al., 2016), and ecological interactions between *Piper* and ants (Wisniewski et al., 2019). A recent study on Aristolochiaceae and other host plants of butterflies (Allio et al., 2021) suggests that the evolutionary success of insects may be linked to recurrent changes in host plants (food sources); these changes have left traces of genetic adaptations in their genomes and are also associated with accelerated diversification. From a morphological point of view, Lactoridaceae, endemic to the Juan Fernández Islands, are unique in angiosperms for their saccate pollen (Zavada and Taylor, 1986). Also unique are the Hydnoraceae, to date the only confirmed family of holoparasitic plants outside the eudicot and monocot radiation, whose type genus was first described as a fungus (Thunberg, 1775). Their extremely modified morphology, including the complete absence of leaves, led Musselman and Visser (1986) to suggest that *Hydnora* is the strangest plant in the world (Thorogood, 2019).

Following the Piperales classification used by Horner et al. (2015), who recognized six families with distinctive morphology, all of which previous studies had recovered as monophyletic, the order consists of: Piperaceae (*Piper*, *Peperomia*, *Manekia*, *Verhuellia,* and *Zippelia*), Saururaceae (*Anemopsis*, *Gymnotheca*, *Houttuynia*, and *Saururus*), Asaraceae (*Asarum*, *Saruma*), Lactoridaceae (*Lactoris*), Hydnoraceae (*Hydnora*, *Prosopanche*), and Aristolochiaceae (*Aristolochia*, *Thottea*). The former two families are the perianth-less Piperales and the latter four are the perianth-bearing members of the order (**Figure 1**). Relationships at the genus level within Piperaceae and Saururaceae are generally well resolved (Meng et al., 2002; Jaramillo et al., 2004; Wanke et al., 2007b; Massoni et al., 2014), unlike those within the perianth-bearing clade. All six family names were validly published in the 18th and 19th centuries, the youngest one more than 130 years ago, and so they have been accepted as well-defined families for a long time (Jussieu, 1789; Giseke, 1792; Ventenat, 1799; Agardh, 1821; Lestiboudois, 1826; Engler, 1887), with generally few changes of taxonomic rank.

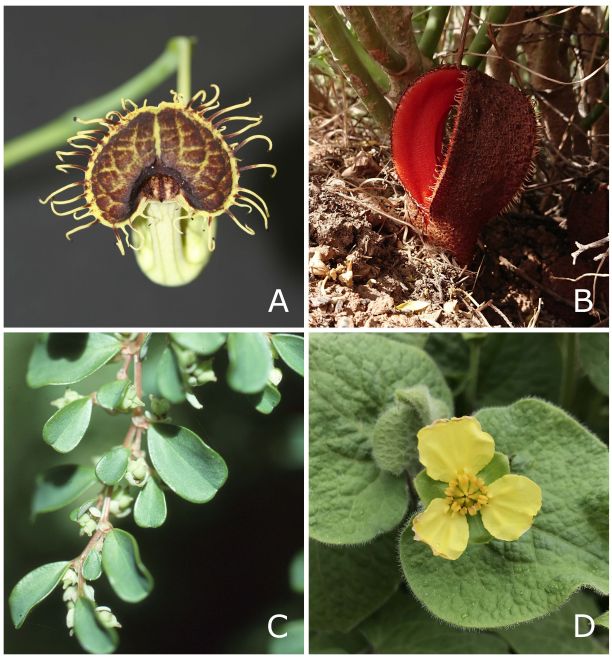


FIGURE 1 | Representatives of perianth-bearing Piperales. **(A)** Flower of *Aristolochia fimbriata* (Aristolochiaceae), **(B)** flower of *Hydnora africana* (Hydnoraceae), **(C)** *Lactoris fernandeziana* with fruits (Lactoridaceae, provided by Tod Stuessy), and **(D)** flower of *Saruma henryi* (Asaraceae, provided by Christoph Neinhuis).

Results of molecular phylogenetic analyses of Piperales in previous studies are generally congruent with respect to the placement of Lactoridaceae. Ignoring the placement of Hydnoraceae, Lactoridaceae are typically recovered as the sister group of Aristolochiaceae, including the studies by Soltis et al. (2000) (based on one mitochondrial and two plastid loci, *Thottea* not included), Qiu et al. (2000) (one nuclear, two mitochondrial, and two plastid loci), Neinhuis et al. (2005) (one plastid locus), Wanke et al. (2007b) (two plastid loci), Wanke et al. (2007a) (one plastid locus), and Massoni et al. (2014) (two nuclear, four mitochondrial, and six plastid loci), and all with poor to moderate support, and with Asaraceae then recovered as the sister group to this clade. However, Jaramillo et al. (2004) recovered a poorly supported clade of Lactoridaceae and Asaraceae, with Aristolochiaceae sister to this clade (one nuclear and two plastid loci), although *Thottea* was missing in their sampling.

Studies that included the holoparasitic Hydnoraceae led to the recovery of multiple different topologies within the perianth-bearing Piperales (Nickrent et al., 2002; Naumann et al., 2013; Massoni et al., 2014). For example, a five-gene analysis (one nuclear, two mitochondrial and two plastid loci) by Nickrent et al. (2002) recovered Hydnoraceae within the

clade of perianth-bearing Piperales, although with poor support, and the whole clade as a polytomy comprising *Aristolochia*, *Lactoris*, a clade of *Asarum* and *Saruma*, as well as a clade of *Hydnora* and *Prosopanche* (the two genera in Hydnoraceae). A six-gene analysis (two nuclear and four plastid loci) in the same study placed Lactoridaceae as the sister group of Hydnoraceae, with Aristolochiaceae then sister to this clade, again with poor support (*Thottea* and Asaraceae were not included in the sampling). Note that in the study by Nickrent et al. (2002), the sampled plastid genes in that study were coded as missing for Hydnoraceae and were later shown to be missing from their plastomes (Naumann et al., 2016; Jost et al., 2020). Naumann et al. (2013) recovered Hydnoraceae as the sister group of Aristolochiaceae from analysis of their 19-gene matrix (14 nuclear, two mitochondrial, and three plastid loci), of which 16 loci are present in Hydnoraceae (although none of the plastid genes). The latter topology had moderate support, with Lactoridaceae sister to the clade comprising Hydnoraceae and Aristolochiaceae. A study that examined 12 loci (two nuclear, four mitochondrial, and six plastid loci) (Massoni et al., 2014) instead recovered Hydnoraceae as the sister group of a clade comprising Lactoridaceae and Aristolochiaceae. In that study, the placement of *Lactoris* as the sister group of *Aristolochia* and *Thottea* received poor support in the maximum likelihood (ML) analysis, as did the sister relationship of *Hydnora* to this clade. *Prosopanche* was not included in their study. The very short estimated branches separating the families in perianth-bearing Piperales are noticeable, and are in close proximity to the extremely long branch leading to *Hydnora*. To date, there has been no phylogenetic study that includes all genera of Piperales.

Apart from these uncertainties on the relationships within the order, the composition of Piperales in terms of its constituent families has also fluctuated in recent angiosperm-wide classification schemes. For example, Angiosperm Phylogeny Group (APG) et al. (2016) accepted only three families in Piperales (Aristolochiaceae, Piperaceae, and Saururaceae), as they decided to lump the families Hydnoraceae and Lactoridaceae with Aristolochiaceae; however, all three families had been recognized in previous iterations of the angiosperm system (APG, 1998, 2003, 2009). APG IV made this decision based on a survey to experts in angiosperm taxonomy addressing various aspects of classification (Christenhusz et al., 2015). However, the question posed to taxonomic experts focused heavily on the position of *Lactoris* in the order, without consideration of Hydnoraceae. Only a single expert noted the phylogenetic evidence on the placement of Hydnoraceae at that time. Despite this, Christenhusz et al. (2015) argued that this did not matter, as Hydnoraceae might also be nested in Aristolochiaceae, and so proposed that it should comprise four subfamilies (i.e., Asaroideae, Hydnoroideae, Aristolochioideae, and the newly proposed Lactoridoideae).

At the time of the survey only three studies had sufficiently sampled the aforementioned families, and each recovered contradictory and poorly supported topologies concerning their interrelationships (Nickrent et al., 2002; Naumann et al., 2013; Massoni et al., 2014). Even ignoring the placement of Hydnoraceae, almost half of the respondents did not favor a three-family system for the order (i.e., ∼46% of experts who voiced their opinion were split between two alternative fragmentations of Aristolochiaceae, biasing the answer to the simpler system). For these reasons, we argue that the suggestions made by Christenhusz et al. (2015) and their implementation in and their implementation in and their implementation in APG (2016) potentially problematic and warrant reconsideration.

Prior to the inclusion of Lactoridaceae and Hydnoraceae in Aristolochiaceae, various studies based on molecular data reported Aristolochiaceae as non-monophyletic, with Lactoridaceae depicted as the sister group of subfamily Aristolochioideae (Qiu et al., 2000; Soltis et al., 2000; Neinhuis et al., 2005; Wanke et al., 2007a,b). In contrast, inferences based on morphological data supported the monophyly of Aristolochiaceae, but were ambiguous about the placement of *Lactoris* (Kelly and González, 2003). The two subfamilies Aristolochioideae and Asaroideae were each recovered as monophyletic in all of these studies. When one traditionally recognized family is placed within another in phylogenetic analyses, Smith et al. (2006) lay out three different options: (1) recognition of the paraphyletic taxon; (2) splitting up the larger family into one or more smaller ones; and (3) lumping the paraphyly-causing family into the family it is nested within. Most systematists, including us, would consider the first option undesirable, but several criteria can be used to decide between the latter two.

One consideration when deciding whether to lump a particular family into another is whether monotypic families should be avoided or not. According to Backlund and Bremer (1998), there is no definitive answer to this question, and arguments for both points of view have to be evaluated based on taxonomic utility. Apart from the primary principle of monophyly following Hennig (1966), Backlund and Bremer proposed secondary principles of classification such as maximizing stability, considering the support for monophyly, the ease of identification, and minimizing redundancy (i.e., maximizing phylogenetic information). These principles are generally followed by APG et al. (2016). Stevens [pers. comm. in Nickrent et al. (2010)] postulates two related principles: the preservation of groups well-established in literature and family size optimization. Additionally, Backlund and Bremer (1998) "*. . .believe that important phylogenetic information is best conveyed by names at the commonly used ranks of genus, family, order. . .*."

Here we present extensive phylogenetic tree inferences for relationships among the genera of Piperales, based on parsimony, likelihood and Bayesian inference (BI) methods, using data from all three subcellular genomes. We then test for potential phylogenomic discordance of inferences based on different genomic compartments, analyze and compare the topological results of the largest sampling of loci for Piperales to date, and conduct several topology tests to evaluate the recovered topologies. Finally, considering our phylogenomic results in perianth-bearing Piperales we discuss arguments for the reconsideration of their classification in light of the principles

described by Hennig (1966), Backlund and Bremer (1998), and Smith et al. (2006).

MATERIALS AND METHODS

Plant Material, DNA Extraction and Sequencing

Fresh leaf material of *Zippelia begoniifolia*, *Manekia incurva*, *Peperomia griseoargentea*, *Verhuellia lunaria*, *Anemopsis californica*, *Gymnotheca chinensis*, *Houttuynia cordata*, *Thottea sumatrana*, and *Saururus cernuus* was collected at the Botanical Garden in Dresden, Germany, cut into smaller fragments and dried in silica gel. Genomic DNA was extracted using the protocol of Doyle and Doyle (1987), modified to include an RNAse A (Thermo Scientific, Waltham, MA, United States) treatment (10 mg/ml). DNA concentration and quality were measured using a Qubit 3 Fluorometer (Thermofisher Scientific, Waltham, MA, United States) and Agilent Technologies 12-capillary Fragment Analyzer (Agilent, 2020) using the genomic DNA 50 kb kit. A paired-end (PE), 300 bp (base pairs) sequencing approach was carried out on a MiSeq (v.3, Illumina, San Diego, CA, United States) with 600 cycles. DNAs were sheared with an M220 ultrasonicator (Covaris, Inc., Woburn, MA, United States) to ∼600 bp and sequencing targeted about five million reads per sample. For *Thottea sumatrana*, ∼4 M. 150 bp PE reads were sequenced on an Illumina NextSeq500 platform with 500 bp insert size. Genome skimming data of *Lactoris fernandeziana* was created based on material used by Graham and Olmstead (2000). Library preparation and size selection followed methods described in Lam et al. (2015). The library was sequenced as 100 bp PE on a HiSeq platform (Illumina, San Diego, CA, United States) to produce ∼6 M. reads. Additionally, unpublished, assembled data for *Hydnora visseri* were provided by Naumann et al. (2016), for *Prosopanche americana* by Jost et al. (2020) and data for *Aristolochia fimbriata* were supplied by Yuannian Jiao (Chinese Academy of Sciences, China) as part of a yet unpublished paper.

Data Mining From Public Repositories to Expand Sampling

Publicly available repositories such as GenBank (NCBI, 2020) and the sequence read archive (SRA, 2020) were mined for assembled organellar genomes or sequencing raw reads with the aim of retrieving data for missing ingroup genera. Additionally, data for one representative for each of several outgroup orders (Amborellales, Nymphaeales, Austrobaileyales, Chloranthales, Magnoliales, Laurales, and Canellales) were extracted to finalize the taxon sampling (**Supplementary Table 1**). Due to the non-availability of data for all three subcellular genomes for a single accession in Canellales, the data of *Drimys* (plastid and mitochondrial) and *Canella* (nuclear) were merged for the concatenated analyses. We are not trying to resolve phylogenetic relationships within the outgroup orders, therefore, this merging is not expected to have an impact on the ingroup results, given that the Canellales terminal serves to anchor that order.

Raw Data Assembly and Extraction of Loci

Raw read data were assembled using the *de novo* assembly function in CLC Genomics Workbench (Qiagen, 2020), allowing for automatic calculation of optimal word and bubble sizes. Gene sequences of all three subcellular genomes for previously published taxa were filtered for the loci of interest (**Supplementary Table 1**). Assemblies were imported into Geneious v.11.1.5 (Geneious, 2020) and individually blasted (BLASTn, evalue 1e-10) for loci of interest from the plastid (pt) and mitochondrial (mt) genomes, using closely related reference species. 83 plastid genes were extracted, consisting of 79 protein coding genes and four ribosomal RNAs (rRNA), 44 mitochondrial genes (41 protein coding and three rRNAs). We also assembled a set of 13 nuclear (nc) loci that are expected to be single or low copy number based on studies of Duarte et al. (2010) and Jiao et al. (2011); those newly sequenced taxa were extracted using a dataset of cDNA sequences by Naumann et al. (2013), while the sampling was expanded with taxa that were obtained from multiple sources and accessions (**Supplementary Table 1**). We aimed for as few sampling gaps as possible; three of originally 13 nuclear loci were excluded from the analyses due to high amounts of missing data (i.e., <50% of sampled species represented).

Phylogenetic Analyses

Single gene alignments were created in Geneious v.11.1.5 (Biomatters, Ltd., New Zealand) using the MAFFT alignment algorithm v.7.450 (Katoh et al., 2002; Katoh and Standley, 2013) and then manually checked and adjusted where necessary in AliView v.1.20 (Larsson, 2014). All genes belonging to the same genome were concatenated with SequenceMatrix v.1.8 (Vaidya et al., 2011), resulting in an 83-gene plastid matrix, a 44-gene mitochondrial matrix and a 10-gene nuclear matrix. A phylogeny based on a maximum likelihood (ML) analysis was created to check and verify that, when different data sources were employed for the same taxon, they were recovered as a monophyletic group when considering each source as a separate operational taxonomic unit (OTU) (**Supplementary Figure 5** MSP), prior to merging them all into the same taxon. In addition to the 83-gene plastid matrix, a 21-gene plastid matrix was created, consisting only of the genes present in either of the two Hydnoraceae genera (Naumann et al., 2016; Jost et al., 2020).

Data were analyzed using parsimony, ML and BI approaches, both per genome and as concatenated sets of plastid, mitochondrial and nuclear data. Parsimony analysis was carried out using PAUP v.4.a165 (Swofford, 1998), implemented in Cipres Science Gateway (Miller et al., 2010) by using 1,000 heuristic searches and 1,000 bootstrap (BS) iterations, with the random starting tree option and the tree bisection-reconnection branch swapping method. Best fitting nucleotide substitution models for different ML analyses were estimated using jModelTest2 v. 2.1.6 (Darriba et al., 2012) and used as input for RAxML v.8.2.12 (Stamatakis, 2014), implemented in Cipres Science Gateway (Miller et al., 2010). ML analysis was carried out on complete data of concatenated gene sets of the individual

genomes. In an attempt to reduce expected long branches leading to Hydnoraceae and to test their overall impact on the topology, we excluded the highly variable third codon position in specific analyses (for protein-coding genes only), and also inferred relationships based on amino acid alignments (protein-coding genes only, translated using Geneious v.11.1.5); although elevated mutational rates in parasitic plants are most apparent in the plastid genome, we repeated these variant analyses for all subcellular genomes, for consistency. The following different data partitioning approaches were also tested to accommodate different patterns of substitution in different subsets of the data: (1) by gene, (2) by gene and codon, (3) by assigning each 3rd codon position its own partition, and (4) unpartitioned (here referred to as single partition). Optimal partitioning schemes in each case were determined using PartitionFinder2 (Lanfear et al., 2014, 2017), and the respective output (i.e., partition combinations and their respective DNA substitution models) were used in the RAxML analyses. For the concatenated plastid, mitochondrial and nuclear data set (137 loci), ML trees were reconstructed using a single partition and a genome partition approach, as well as a translated (single partition) amino-acid sequence alignment. For all ML analyses, 1,000 bootstrap iterations were calculated. Finally, BI tree estimates for the fully concatenated unpartitioned and genome partitioned case were done using MrBayes v.3.2.7a (Huelsenbeck and Ronquist, 2001) on Cipres Science Gateway (Miller et al., 2010) with four chains and calculating 20×10^6 generations, after which chains converged (assessed using the estimated sample size ESS) and a burn-in of 2×10^6 was chosen. In addition to the above analyses, each complete single genome and concatenated nucleotide data set was considered with Hydnoraceae excluded (gene/genome partition, RAxML, 1,000 bootstrap iterations, **Supplementary Figure 1**), and using the genome partition approach we also performed a concatenated analysis of plastid and mitochondrial data only, including Hydnoraceae (RAxML, 1,000 bootstrap iterations, **Supplementary Figure 6**). Lastly, for the nuclear data set a coalescent tree was estimated using ASTRAL v.5.6.3 (Mirarab and Warnow, 2015; Sayyari and Mirarab, 2016), based on single gene trees. All trees were visualized using TreeGraph 2 (Stöver and Müller, 2010), with Amborellales defined as the outgroup. Taxon names in the phylogenetic trees are represented with either a binominal or genus only, depending on whether different accessions for a single genus were merged (in the latter case, sometimes different species, see above) to achieve the best locus-level coverage (**Supplementary Table 1**).

Topology Testing

All 15 different, possible tree topologies for the four main lineages in the monophyletic perianth-bearing Piperales clade were tested for their significance using the tree topology evaluation tests implemented in IQ-Tree (Nguyen et al., 2015). Five of these topologies were recovered in one or more of our phylogenetic tree reconstructions. The tree files for the remaining ten topologies were manually created by altering only the relationships in the clade of interest. The bootstrap proportions using RELL (Kishino et al., 1990), SH test (Shimodaira and Hasegawa, 1999), weighted SH test, expected likelihood weight (ELW, Strimmer and Rambaut, 2002) and the approximately unbiased test (Shimodaira, 2002) were carried out in multiple runs. All tests performed 10,000 resamplings using the RELL method. We carried out five independent runs, one for each different topology recovered in our analyses. The program was provided with both the alignment file and substitution model used to infer the best tree for that data set (null hypothesis), as well as the set of alternative hypotheses (tree file containing all 15 possible topologies). For example, run A (**Figure 4A**) was provided with the data set reconstructing topology 1, and run B (**Figure 4B**) was provided with the data set underlying topology 2; significance was evaluated considering all 15 topologies.

RESULTS

Dataset Characteristics

Assembly of newly generated next generation sequencing (NGS) data and database-mined loci of interest recovered varying amounts of data per accession and genome (**Table 1** and **Supplementary Table 1**). 83 plastid loci were recovered for nearly all taxa sampled, with the exception of Hydnoraceae whose two genera have plastomes greatly reduced in gene content: here, only 20 plastid genes could be used for phylogenetic tree reconstruction. A total of 44 mitochondrial markers were recovered for almost all newly sequenced accessions, but fewer loci were retrieved for certain taxa sampled from GenBank (e.g., less than 50% of the total mt loci could be mined for *Asarum*, *Chloranthus*, *Drimys*, and *Saruma*). With regard to the recovered number of loci and the overall locus coverage, the nuclear data set was the most variable (**Supplementary Figure 7**). Complete coverage of all ten nuclear loci was achieved for only three accessions (*Aristolochia*, *Liriodendron*, and *Piper*), with only a single locus available for *Schisandra* and *Prosopanche* (**Table 1** and **Supplementary Figure 7**).

Molecular Phylogenomic Tree Reconstruction

Phylogenetic Tree Reconstructions Excluding Hydnoraceae

When Hydnoraceae are excluded from the datasets, virtually identical relationships are recovered across all analyses. Within the perianthless Piperales, Saururaceae, and Piperaceae are reconstructed as monophyletic and branch support values are very high (**Supplementary Figure 1**). In the latter family, *Manekia* + *Zippelia* is sister to *Peperomia* + *Piper,* and *Verhuellia* is sister to this entire clade. Within Saururaceae, the clades comprising *Gymnotheca* + *Saururus*, and *Anemopsis* + *Houttuynia*, have 100% support in all analyses, except in the analysis of nuclear data alone (**Supplementary Figure 1**). In the latter, *Anemopsis* is sister to the clade of *Gymnotheca* + *Saururus* with low support (BS 47%) and *Houttuynia* sister to the entire clade (BS 100%). Within perianth-bearing Piperales, relationships are identical between the plastid, mitochondrial, and concatenated data-based analyses (**Supplementary Figure 1**). Asaraceae are sister to the clade

TABLE 1 | Overview of the number of character sets (charsets) and total sequence length (bp) recovered for the three subcellular genomes for all individual accessions represented in the sampling (see **Supplementary Table 1** for more details).

Taxon	Plastid		Mitochondrial		Nuclear	
	No. of charsets	Total length (bp)	No. of charsets	Total length (bp)	No. of charsets	Total length (bp)
Amborella	83	86,218	39	45,796	9	5,186 bp
Anemopsis	83	86,222	43	46,972	8	2,143 bp
Aristolochia	83	86,222	44	50,050	10	6,422 bp
Asarum	82	85,466	20	22,110	6	2,716 bp
Calycanthus	83	86,195	44	45,399	9	4,764 bp
Chloranthus	83	86,211	10	19,230	6	3,981 bp
Drimys/Canella	83	86,212	11	19,714	9	2,176 bp
Gymnotheca	83	79,999	44	48,831	6	1,821 bp
Houttuynia	83	86,222	42	47,333	9	3,170 bp
Hydnora	20	34,324	42	47,062	9	5,192 bp
Lactoris	83	86,213	28	36,744	7	2,662 bp
Liriodendron	82	85,285	44	50,563	10	6,104 bp
Manekia	83	86,222	44	49,414	9	4,232 bp
Nymphaea	83	86,222	42	43,791	9	5,979 bp
Peperomia	83	86,222	41	41,551	5	2,500 bp
Piper	83	86,215	44	48,415	10	6,440 bp
Prosopanche	20	34,688	42	47,360	1	345 bp
Saruma	83	86,215	7	15,569	9	4,953 bp
Saururus	82	85,865	42	47,597	7	2,839 bp
Schisandra	83	86,195	44	50,233	1	482 bp
Thottea	83	86,222	44	50,010	9	3,374 bp
Verhuellia	83	86,222	44	50,446	8	3,161 bp
Zippelia	83	86,222	44	49,620	5	1,019 bp

Accessions are ordered alphabetically. A maximum of 83 plastid loci, 44 mitochondrial loci, and 10 nuclear loci was aimed for. Taxa Drimys *and* Canella *have been merged to achieve coverage for all three subcellular genomes for outgroup Canellales.*

of Aristolochiaceae and Lactoridaceae with moderate to full support (BS 79–100%) and all families are monophyletic. Tree reconstruction based on the concatenated 10 nuclear loci recovered the clade Lactoridaceae and Asaraceae (BS 56%) sister to Aristolochiaceae + monophyletic perianthless Piperales (BS 83%, **Supplementary Figure 1**). Branch lengths within trees are relatively homogenous, with the shortest branches in Piperales recovered across the four data sets within Saururaceae, as well as within perianth-bearing Piperales and the branch leading to the latter (**Supplementary Figure 1**).

Hereafter, we only describe in detail the relationships within perianth-bearing Piperales; relationships within the perianth-less clade can be found, for each analysis, in the supporting material (**Supplementary Figures 3–6**). The topology within the latter clade is consistent across data sets, as well as types of analysis with very strong support, with the exception of some analyses based on nuclear data alone (**Supplementary Figure 5**).

Phylogenetic Tree Reconstructions Including Hydnoraceae

Inclusion of Hydnoraceae leads to varying topologies depending on the subcellular origin of the data. Concatenated data or data with organellar origin typically recover two topologies for relationships within perianth-bearing Piperales, differing only in the relationships within the clade consisting of Aristolochiaceae, Hydnoraceae, and Lactoridaceae (**Table 2** and **Figures 2A,B**). First, this three-family clade (referred to as clade I in **Table 2**) is well-supported in most analyses, although more weakly supported by the various analyses involving plastid data alone. Within this clade, a sub-clade comprising Aristolochiaceae and Lactoridaceae (referred to as clade II in **Table 2**), is recovered with poor to strong support for five out of seven analyses based on the concatenated organellar and concatenated three-genome data set (BS 24–79%, PP ∼0.99, **Supplementary Figure 6** and **Table 2**) and with moderate support for all of the plastid-data derived inferences (BS 51–87%, PP 0.61–0.75; **Figure 2A** and **Supplementary Figure 3**); it tends to be less well-supported by partitioned data and is best supported in the 21-gene analysis of plastid data alone (BS 89%). The latter analysis includes only the plastid genes present in either of the Hydnoraceae genera. In contrast, a sub-clade comprising Aristolochiaceae and Hydnoraceae (referred to as clade III in **Table 2** and **Figure 2B**) is recovered in all tree reconstructions based on mitochondrial data alone (ML, MP, and BI), with weak to strong support for this relationship (BS 56–88%, PP 0.96–1), and generally better support in partitioned likelihood analyses than the unpartitioned one. Clade III is also recovered in some inferences based on nuclear data alone (**Figure 2C** and **Supplementary Figure 5**), with low to moderate support (BS 23–75%, PP 0.96–0.97). Deletion of the third codon position appears to have little effect

TABLE 2 | Summary of bootstrap and posterior probability support for analyses supporting the two predominantly recovered topologies within perianth-bearing Piperales.

	All combined				Organellar only		Plastid							
	137 SP	137 GnP	137 SP	137 GnP	127 SP	127 GnP	83 SP	83 GP	83 GCP	83 3SP	83 3GP	83 SP	83 GP	21 SP
Clade	**ML**	**ML**	**BI**	**BI**	**ML**	**ML**	**ML**	**ML**	**ML**	**ML**	**ML**	**BI**	**BI**	**ML**
I	99	100	1	1	98	100	44	69	68	–	50	–	0.581	50
II	36	79	0.996	0.999	24	72	87	59	51	85	68	0.753	0.606	89

	Mitochondrial								
	44 SP	44GP	44 GCP	44 3SP	44 3GP	44 ASP	44 SP	44 SP	44 GP
	ML	**ML**	**ML**	**ML**	**ML**	**ML**	**MP**	**BI**	**BI**
I	86	93	95	81	83	81	74	1	1
III	64	88	84	56	81	57	62	0.964	1

Clade I refers to the clade comprising Aristolochiaceae, Hydnoraceae, and Lactoridaceae, clade II to the sister relationship of Aristolochiaceae and Lactoridaceae, and clade III to the sister relationship of Aristolochiaceae and Hydnoraceae. Analyses are displayed as number of loci and partitioning approach used, with single partition (SP), genome Partition (GnP), gene partition (GP), gene by codon partition (GCP), exclusion of the 3rd codon position (3SP, 3GP), and translated amino-acid data (ASP). Bootstrap support values are displayed for ML analyses and posterior probability for BI analyses. Dashes (–) highlight analyses without representation of a certain clade. Visual representation of the clades can be found in ***Figures 2****,* ***3****, as well as* ***Supplementary Figures 3, 4, 6****.*

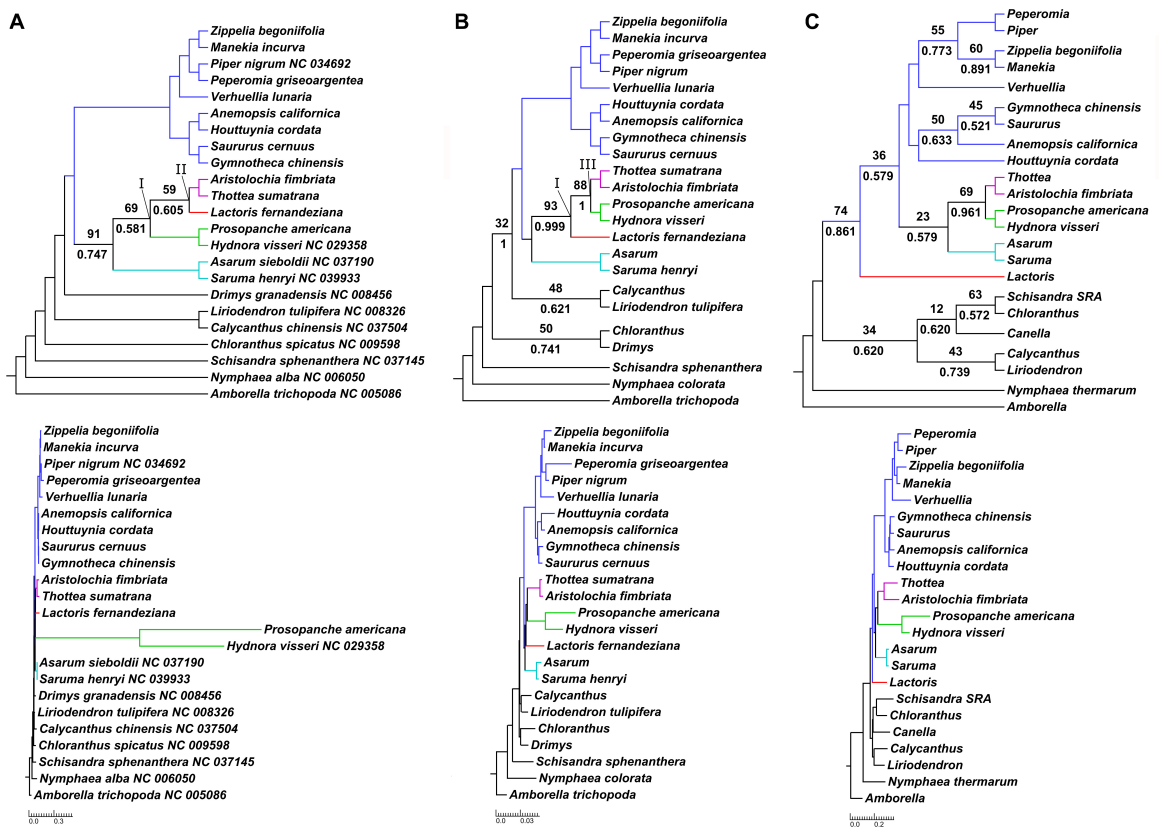


FIGURE 2 | Phylogenomic discordance between plastid, mitochondrial and nuclear datasets. Results of maximum likelihood and Bayesian inference tree reconstruction based on gene-partitioned nucleotide data, as a cladogram with corresponding phylogram below, for **(A)** 83 plastid loci; **(B)** 44 mitochondrial loci; and **(C)** 10 nuclear loci. Support values are displayed for branches with <95% bootstrap support (above branch) or <0.95 posterior probability (below branch). Bootstrap support values are based on 1,000 pseudoreplicates. Piperales are color-coded: Asaraceae in turquoise, Aristolochiaceae in purple, Lactoridaceae in red and Hydnoraceae in green, Piperaceae and Saururaceae in blue. The three-family clade Aristolochiaceae, Hydnoraceae and Lactoridaceae is annotated with I, the clade Aristolochiaceae + Lactoridaceae with II, and the clade Aristolochiaceae + Hydnoraceae with III.

on support for either clade II or III for plastid or mitochondrial data (**Table 2**).

Phylogenetic Tree Reconstructions Recovering Additional Topologies

Parsimony analysis of the concatenated 137-loci set recovers Lactoridaceae sister to Hydnoraceae, and Aristolochiaceae sister to that clade (**Supplementary Figure 6**). Tree reconstruction based on the amino-acid alignment of the same data set places Lactoridaceae sister to Aristolochiaceae + Hydnoraceae (**Supplementary Figure 6**), although with low support for the clade Aristolochiaceae + Hydnoraceae (BS 51%). Based on plastid data alone, Hydnoraceae were twice estimated to be sister to all remaining perianth-bearing Piperales although with poor support (BS 52–54%) (**Supplementary Figure 3** ML, CP, and 3SP). A placement of Hydnoraceae close to the root of angiosperms was estimated for the translated amino-acid plastid data (Austrobaileyales sister to Hydnoraceae, **Supplementary Figure 3** ASP) and MP analysis of the nucleotide data (Nymphaeales sister to Hydnoraceae, **Supplementary Figure 3**). Nuclear data-based tree reconstruction recovers Hydnoraceae sister to Aristolochiaceae in nine out of ten analyses (**Figure 2C** and **Supplementary Figure 5**) with weak to moderate support in ML analyses (BS 23–75%) and strong support in BI (PP 0.96–0.97). Asaraceae are placed sister to the aforementioned clade in multiple analyses (e.g., **Figure 2C**). Lactoridaceae placement is mostly ambiguous and poorly supported with either Lactoridaceae sister to Asaraceae (e.g., **Supplementary Figure 5** ML and 10 SP) or sister to all other Piperales (e.g., **Figure 2C**). The inference based on coalescent analysis of the ten nuclear loci differed in some cases drastically from the concatenated one, based on the same input data set (**Supplementary Figure 5**). The coalescent analysis recovers paraphyletic perianth-bearing Piperales with Aristolochiaceae sister to the perianth-less clade and with Lactoridaceae sister to the clade Asaraceae + Hydnoraceae. Analyses with the third codon position excluded or based on an amino-acid alignment recover the paraphyly of perianth-bearing Piperales (**Supplementary Figure 5**), and the latter analysis also recovers the paraphyly of Hydnoraceae, although with poor support. Parsimony and BI recover a large polytomy, sometimes including multiple outgroup taxa (**Supplementary Figure 5**).

Topology Testing

Within the perianth-bearing Piperales, a total of five discordant topologies with this clade monophyletic are recovered in this study and tested alongside the other 10 possible ones for their significance (**Figure 4** and **Supplementary Figure 2**; there are 15 possible rooted arrangements of the four families, shown at the foot of **Supplementary Figure 2**, and note that the first five topologies in the latter are in the same order as the former figure). In summary, the first topology (**Figures 3**, **4.1**) is recovered from the 137-loci combined analysis (organellar + nuclear data), using ML and partitioning by genome. The second topology (**Supplementary Figure 3** ML 83, CP and **Figure 4.2**) was estimated using the 83 plastid data set and assigning the 3rd codon position its own partition. The topology reconstructed using the mitochondrial data set (ML analysis and gene partition, **Figures 2B**, **4.3**) is the third topology tested. The fourth one is the result of the maximum parsimony analysis of the 137-loci data set (**Supplementary Figure 6** and **Figure 4.4**), and the fifth topology was estimated for the ML analysis of the concatenated nuclear loci (single partition, **Supplementary Figure 5** and **Figure 4.5**). In total, five independent analyses were run to test whether a specific data set rejects a certain topology. All runs were provided the identical set of topologies, corresponding to all possible topologies for a monophyletic perianth-bearing Piperales clade. The topologies differ only in the inferred relationships within the perianth-bearing Piperales (**Figures 4.1–5**). The runs themselves differed in the data set chosen as null hypothesis, e.g., run A (**Figure 4A**) was provided with the data set reconstructing topology 1, and run B (**Figure 4B**) was provided with the data set underlying topology 2.

All topologies performed best when the underlying data were set as null hypothesis, with the exception of topology 4, which performed only second best behind topology 1 (**Figure 4D**), although the null topology here was recovered using parsimony, not likelihood. Topology 1 (Hydnoraceae sister to the clade of Aristolochiaceae + Lactoridaceae) performed best in two out of five runs and was only significantly excluded twice (with exception of SH and WSH in run C and SH of run E). Both topology 2 (Hydnoraceae sister to all other perianth-bearing Piperales) and topology 3 (Lactoridaceae sister to the clade of Aristolochiaceae + Hydnoraceae) were significantly excluded in four out of five runs, except when their underlying data were set as the null hypothesis. Topology 4 (Aristolochiaceae sister to the clade of Hydnoraceae + Lactoridaceae) was only significantly excluded in the RELL and AU tests with topology 1 set as null hypothesis (**Figure 4A**), but was in no run the best performing one. Lastly, topology 5 (Lactoridaceae sister to Asaraceae and this clade sister to Aristolochiaceae + Hydnoraceae) was rejected by all analyses, except when it was set as the null hypothesis (**Figure 4E**). Run E also rejected all other tested topologies (with exception of the SH test).

The majority of the additional ten topologies (not recovered in this study; trees 6–10 in **Supplementary Figure 2**) were rejected by all tests in runs A, B, and D. One exception being topology 6 (Lactoridaceae + Aristolochiaceae sister to the clade of Asaraceae + Hydnoraceae) in run B (**Supplementary Figure 2**). Not rejected, but poorly performing are many of the additional topologies for run C, as well as the SH test of run E (**Supplementary Figure 2**). Topology 11 (Asaraceae sister to the clade of Aristolochiaceae + Hydnoraceae, with Lactoridaceae sister to this whole clade) is the only of the ten topologies rejected in only three out of five runs, yet not recovered in any of our inferences.

DISCUSSION

Phylogenomic Discordance Among Genomic Compartments

Phylogenetic tree reconstructions of the magnoliid order Piperales at the genus level, excluding holoparasitic Hydnoraceae,

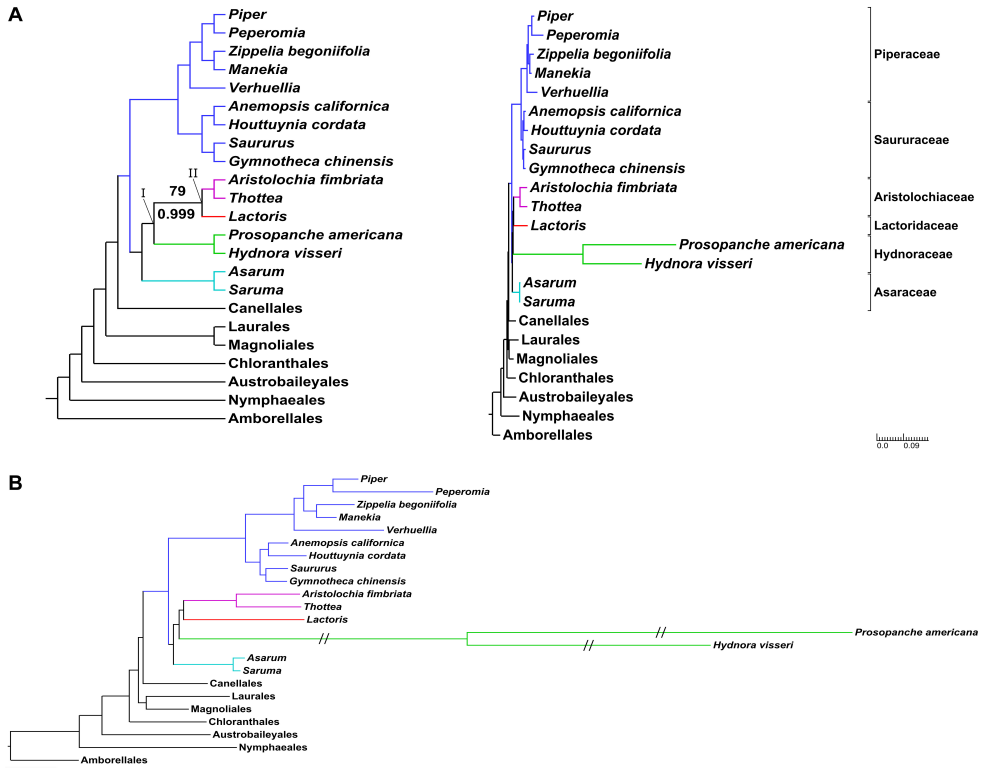
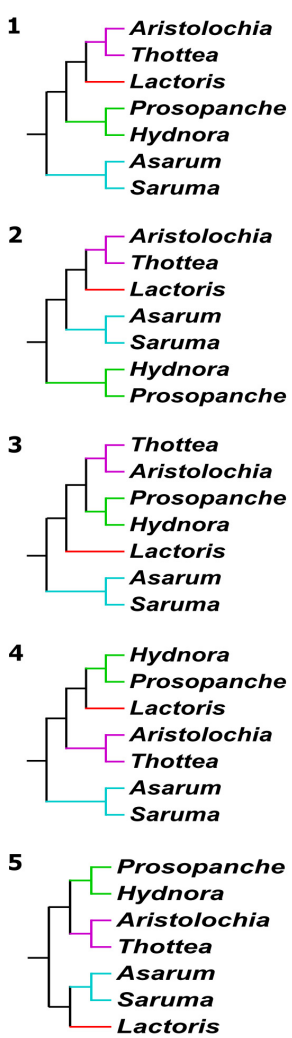


FIGURE 3 | Piperales phylogeny based on genome-partitioned 137 loci of three subcellular genomes. **(A)** Topology with full support (BS 100%, PP 1) for all nodes, unless indicated for maximum likelihood analysis (bootstrap, above the branch) and Bayesian inference (posterior probability, below the branch). The right side shows the phylogram with annotated families of the Piperales. **(B)** Amplified view of branch lengths within Piperales. Branches annotated with "//" are shortened by 50%. Piperales are color-coded: Asaraceae in turquoise, Aristolochiaceae in purple, Lactoridaceae in red and Hydnoraceae in green, Piperaceae and Saururaceae in blue. The three-family clade Aristolochiaceae, Hydnoraceae and Lactoridaceae is annotated with I, the clade Aristolochiaceae + Lactoridaceae with II.

recover the perianth-less Piperales clade; both inferences based on organellar data also recover a perianth-bearing clade (**Supplementary Figure 1**). The former comprises the two monophyletic families Piperaceae and Saururaceae. Within perianth-bearing Piperales, three clades are recovered, with Asaraceae sister to a clade of Lactoridaceae + Aristolochiaceae; these relationships received strong bootstrap support (BS 99–100%) for the concatenated and plastid data sets, as did those within perianthless Piperales based on mitochondrial data (BS 100%). Taxon bipartitions within perianth-bearing Piperales are well-supported based on the latter data. Nuclear-based phylogenies also recovered both Piperaceae and Saururaceae as monophyletic, although with lower support than in the aforementioned data sets, and perianth-bearing Piperales are recovered as non-monophyletic with weak support. The analyses based on nuclear single-copy locus data are potentially biased by the amount of missing data for several accessions (**Table 1** and **Supplementary Figure 7**). Although this nuclear result could be based on cytonuclear discordance (e.g., shown in asterids, Stull et al., 2020) between Lactoridaceae and the other members of the perianth-bearing Piperales, we also cannot rule out the possibility that undiagnosed paralogy in subsets of the nuclear loci, particularly given the mixed sources of data for this subcellular genome (a combination of Sanger sequencing, some mined data without read information, and genome skimming with lower coverage for these loci). For example, in Rosaceae it has recently been shown, that for many "single copy" loci used in common target enrichment, paralogs can be found with increasing sequencing depth reflecting ancient gene duplication (Morales-Briones et al., 2020).

Extensive phylogenetic tree reconstructions that include the holoparasitic Hydnoraceae predominantly recover two topologies across data sets—differing only in the relationships within the clade comprising Aristolochiaceae, Hydnoraceae, and Lactoridaceae. The case with Hydnoraceae sister to Aristolochiaceae + Lactoridaceae (clade II) was generally well-supported for inferences based on loci from all three subcellular genomes combined (**Figure 3**), concatenated, organellar genomes only (**Supplementary Figure 6**, **127** OSP/OGnP) and plastid data alone (**Figure 2A** and **Supplementary Figure 3**). Support for this clade is highest when all data gathered were analyzed together, with moderate to strong support (both for ML and BI). The sister relationship of Hydnoraceae to the clade Lactoridaceae + Aristolochiaceae (e.g., **Figure 3**) is identical to the one recovered by Massoni et al. (2014), although here with the inclusion of *Prosopanche* and branches being well supported, for both the ML and BI analyses. Similar to previous studies (Massoni et al., 2014; Wanke et al., 2017), short branches,



A

Tree	bp-RELL	p-SH	p-WSH	c-ELW	p-AU
1	0.958	1	1	0.941	0.973
2	0.0013	0.7	0.0159	0.00161	0.00106
3	0	0.0003	0	4.21e-77	1.53e-05
4	0.0398	0.855	0.204	0.056	0.0424
5	0	0	0	0	2.75e-05

B

Tree	bp-RELL	p-SH	p-WSH	c-ELW	p-AU
1	0.187	0.994	0.98	0.267	0.718
2	0.279	1	0.974	0.268	0.529
3	0	0.0005	0	9.63e-63	7,00E-55
4	0.172	0.897	0.603	0.161	0.201
5	0	0	0	0	5.87e-10

C

Tree	bp-RELL	p-SH	p-WSH	c-ELW	p-AU
1	0.0006	0.235	0.211	0.00063	0.00624
2	0	0.0553	0.0138	2.17e-16	3.6e-53
3	0.65	1	0.969	0.648	0.801
4	0.154	0.631	0.706	0.155	0.355
5	0	0	0	2.23e-144	0.061

D

Tree	bp-RELL	p-SH	p-WSH	c-ELW	p-AU
1	0.933	1	1	0.917	0.97
2	0.0083	0.711	0.0308	0.00379	0.00297
3	0	0.0002	0	1.44e-67	2.42e-24
4	0.0629	0.859	0.267	0.076	0.0478
5	0	0	0	0	3.3e-31

E

Tree	bp-RELL	p-SH	p-WSH	c-ELW	p-AU
1	0.0001	0.476	0.0155	0.000274	0.00331
2	0	0.463	0.0048	3.59e-06	0.00548
3	0.0072	0.497	0.0496	0.00737	0.015
4	0	0	0	0	3.92e-05
5	0.897	1	1	0.885	0.984

FIGURE 4 | Topology test results for recovered topologies. Shown are the tested topologies **(1–5)** and tables **(A–E)** containing results of the bp-RELL, p-SH, p-WSH, c-ELW, and p-AU analyses. Tree numbers in the tables correspond to the topologies on the left **(1–5)**. Topology **(1)** contains clade II and was recovered for the concatenated three-genome data set (ML, genome partition), topology **(2)** was inferred for the 83 gene plastid data set (ML, assigning each 3rd codon position its own partition). Topology **(3)**, containing clade III, was reconstructed using the 44 gene mitochondrial data set (ML, gene partition), topology **(4)** using the concatenated 137-loci data set (MP, single partition), and topology **(5)** using the concatenated nuclear loci (ML, single partition). Table **(A)** shows the topology test results with topology **(1)** as null hypothesis, table **(B)** with topology **(2)** as null hypothesis and so on. Blue-colored values denote results within the 95% confidence sets; red-colored values denote significant exclusion. In the topologies **(1–5)**, perianth-bearing Piperales are color-coded: Asaraceae in turquoise, Aristolochiaceae in purple, Lactoridaceae in red and Hydnoraceae in green.

especially within perianth-bearing Piperales, are situated in close proximity to extremely long branches, not only leading to Hydnoraceae (Massoni et al., 2014), but also to the respective terminal branches for *Prosopanche* and *Hydnora* (**Figure 3B**; Jost et al., 2020). These drastic differences in branch lengths, together with the reduced number of available plastid markers, likely contributed to difficulties in previous studies that attempted to place these holoparasites. Analyses based on mitochondrial loci alone recover a different set of relationships (i.e., clade III instead of clade II; **Figure 2** and **Table 2**). All inferences based on 44 mitochondrial loci recover Lactoridaceae as sister to the clade comprising Aristolochiaceae + Hydnoraceae, with low to strong support for ML, MP, and BI (**Figure 2B** and **Supplementary Figure 4**). In contrast to the plastid and concatenated results, branch lengths for the mitochondrial inferences are more homogenous across the tree, with short branches at the backbone in perianth-bearing Piperales but no drastic increase in Hydnoraceae. Phylogenomic discordance between the two organellar genomes is reflected not only by differences in topology, but also in branch lengths (rates of evolution), arising from a drastically reduced and rapidly evolving Hydnoraceae plastome (Naumann et al., 2016; Jost et al., 2020), in contrast to a mitochondrial genome that is presumably evolving at rates consistent with those of photosynthetic plants. Relationships inferred among outgroup orders also vary between different analyses based on mitochondrial data, which is most

likely a result of the low number of loci derived from GenBank for some accessions (**Table 1**). The greatest differences with respect to number of loci and base pairs of sequence recovered are for the nuclear data, with *Prosopanche* and *Schisandra* represented by only a single locus (**Table 1**). These factors are most likely the reason for the various more unusual topologies recovered when reconstructing relationships using the nuclear data alone. Despite those differences, topologies within the perianth-less Piperales inferred based on the nuclear data are relatively stable, as are those within the perianth-bearing clade, with the placement of Lactoridaceae being the exception. Nonetheless, adding the nuclear data to the concatenated organellar data increases the support in comparison to the organellar data alone (**Supplementary Figure 6**).

The sister relationship of Lactoridaceae + Hydnoraceae, also inferred in the six-gene analysis of Nickrent et al. (2002), was recovered here in the MP analyses of our concatenated 137-loci data set (**Supplementary Figure 6**). Tree inferences in analyses that include plastid loci are potentially negatively affected by long branch attraction (LBA, Felsenstein, 1978; Hendy and Penny, 1989) when using a parsimony approach, and therefore might differ from inferences estimated using model-based methods (ML and BI). The latter phenomenon has previously been confirmed in, for example, holoparasitic Rafflesiales (Nickrent et al., 2004) and mycoheterotrophic plants (Lam et al., 2018). In our study, this is likely the case apparent when comparing likelihood results to the parsimony tree estimation of the mitochondrial data. Here, with mostly homogenous branch lengths across the mitochondrial tree, LBA is less likely to affect placement of taxa.

Overall, inferences based on mitochondrial data alone proved to be the most consistent across analyses with regards to topology. Topologies within Piperales were identical, regardless of analysis type (ML, MP, and BI), data reduction (3rd codon position excluded, translated amino-acid alignment) and partitioning approach. Inferences based on the concatenated three-genome data recovered an identical topology to the predominantly recovered one based on plastid data alone (clade II), though with much higher support for branches within perianth-bearing Piperales. Across all performed analyses, generally the use of gene partitioned ML analyses (genome partition for the concatenated analyses) tended to lead to the highest support values. Removing data subpartitions that are rapidly evolving (the 3rd codon position) or using amino-acid data and amino-acid substitution models (amino-acids evolve slower than nucleotide data) were unsuccessful in the sense that they yielded poorly supported trees with in some cases altered topology (**Supplementary Figure 3** ASP); this may simply be a function of having too little data to make robust inferences in these cases.

The Most Likely Phylogenetic Relationships Within Perianth-Bearing Piperales

Considering all the evidence, the most likely topology for relationships within perianth-bearing Piperales is the one recovered for the concatenated three genome analysis (**Figure 3**), with strong to full BS and PP support for Hydnoraceae sister to Lactoridaceae + Aristolochiaceae, and with Asaraceae sister to that clade. These results are identical to the poorly supported topology reported by Massoni et al. (2014), but here, with both genera of Hydnoraceae included and additional data considered per taxon, these relationships are well-supported. This topology receives additional support from the results of the conducted topology tests, evaluating the significance of all recovered relationships within perianth-bearing Piperales in comparison to one another, as well as in comparison to all other possible (but not recovered) topologies for the four families. The topology recovered by the six-gene analysis of Nickrent et al. (2002) is significantly excluded by the topology testing (in analysis 2 and 3), as well as the topology recovered for all analyses solely based on mitochondrial data (**Figures 2B**, **4.3**, clade III), highlighting the discordance of genetic signals recovered for the two organellar genomes (**Table 2**). Within perianth-bearing Piperales, the uncertain placement may well be attributable to extremely short branches in close proximity to the extremely long ones that lead to Hydnoraceae. Missing plastid markers owing to plastome size reduction (Naumann et al., 2016; Jost et al., 2020), together with limited accessibility of plant material for Lactoridaceae, have made placement of Hydnoraceae difficult to infer in previous studies (Nickrent et al., 2010; Naumann et al., 2013; Massoni et al., 2014).

Thoughts on Classification Within Perianth-Bearing Piperales

The classification of Piperales implemented by APG et al. (2016), prompted by the online survey of Christenhusz et al. (2015), needs reconsideration. Furthermore, discussions prompted by this survey are not only limited to this order. For example, Nyffeler and Eggli (2020) argued that the lumping done by APG within Asparagales "*. . .does not result in a gain of information*" and they argue to instead follow more traditional family circumscriptions until the proposed argument for higher practicability in Christenhusz et al. (2015) is proven. A similar argument was made by Nickrent (2020) against lumping of taxa in Santalales by APG et al. (2016). In Piperales, the lumping of Hydnoraceae and Lactoridaceae into Aristolochiaceae was based on two contradictory topologies available at that time (Naumann et al., 2013; Massoni et al., 2014). We argue that the problem of paraphyly in Aristolochiaceae s.l. (*Aristolochia*, *Asarum*, *Saruma*, and *Thottea*), also demonstrated in previous studies (Qiu et al., 2000; Soltis et al., 2000; Neinhuis et al., 2005; Wanke et al., 2007a,b) cannot simply be swept under the carpet by lumping Hydnoraceae and Lactoridaceae as well. A debate based on phylogenetic evidence, which we present here, has to be held, and the solutions that Smith et al. (2006) propose for such cases also have to be evaluated.

What are the alternatives and how do we decide among them? With the sound placement of Lactoridaceae and Hydnoraceae within Aristolochiaceae s.l., the latter could be recognized as a paraphyletic family, or split into multiple smaller ones, or the former two could be lumped into the family they are nested in, a broadly defined and monophyletic Aristolochiaceae.

The first case (paraphyly) is generally undesirable, and the latter was recommended by Christenhusz et al. (2015) and implemented by APG et al. (2016). Lumping of the three families into Aristolochiaceae reduces Lactoridaceae and Hydnoraceae to subfamily status. While subfamilies Aristolochioideae Link, Asaroideae O. C. Schmidt and Hydnoroideae Walpers were previously described, they are rarely used. In addition, subfamily Lactoridoideae was not validly published by Christenhusz et al. (2015) according to the ICN (Art. 41.5, Turland et al., 2018), as the page number of the publication of its basionym Lactoridaceae was omitted (i.e., T.3 Abt.2: 19, Engler, 1887). To our knowledge this error has not been corrected elsewhere, and both Mabberley (2017) and Stevens (2001 onward) did not use this subfamily name and instead mentioned the genus name *Lactoris* along with the names of the other three subfamilies. There is no advantage in using subfamily over family names when both represent identical clades, especially if one of the subfamilies has to be newly introduced and an already established corresponding family name is available. Therefore, based on our data, we support the recognition of Hydnoraceae and Lactoridaceae, and a reversion to the earlier APG classifications (APG, 2003, 2009). We therefore accept four monophyletic families within the perianth-bearing Piperales, in line with Horner et al. (2015) and Nickrent (2020). Recognition of a narrowly defined Aristolochiaceae also requires recognition of Asaraceae, containing *Asarum* and *Saruma*, which are not closely related to Aristolochiaceae.

This classification with Asaraceae as a recognized family was also proposed by Nickrent (2020) who stated that this system within perianth-bearing Piperales "*. . .would result in the least amount of disruption*" and "*. . .would recognize the morphological distinctions among the members.*" The primary principle of monophyly of Hennig (1966) and the secondary principles of Backlund and Bremer (1998) are also met with our approach, which is maximizing stability, the support for monophyly, and minimizing redundancy. Additionally, each of the four distinct families within perianth-bearing Piperales are supported by clear apomorphies (see e.g., Stevens, 2001, onward), thus "*maximizing the ease of identification*" (Backlund and Bremer, 1998). The additional principle of preservation of groups well-established in the literature (Steven's pers. comm. in Nickrent et al., 2010) is also met. As a service to society, a fundamental aspect of classification is its predictive quality (Stuessy, 2009a,b, 2013). The alternative approach of having broad classifications with fewer families places this aspect at risk, especially for lineages that are relatively unknown to many researchers and the general public (undoubtedly the case with Hydnoraceae and Lactoridaceae). We argue that a better approach is therefore to recognize multiple families of perianthless Piperales.

Additional Considerations on the Families

Lactoridaceae, with its single remaining species *Lactoris fernandeziana*, are a relic of early angiosperm evolution (Stuessy et al., 1998) and are currently found only on a single island of the Juan Fernández Archipelago, Chile. Although the Juan Fernández Islands are relatively young volcanic islands (Stuessy et al., 1984; Ricci, 2001), fossil pollen of *Lactoripollenites* (=*Rosannia*) is widespread in the fossil record, from Late Cretaceous deposits from Namibia (Turonian-Campanian) to India, Australia, and North and South America (Zavada and Benson, 1987; Macphail et al., 1999; Gamerro and Barreda, 2008; Srivastava and Braman, 2010). Lactoridaceae are the only endemic angiosperm family of the Juan Fernández Islands and are an important signature plant for conservational efforts on the island flora. If Lactoridaceae were to lose their family status, this could impact the political acceptance of the conservational efforts (Stuessy et al., 2014). Ideally, political considerations must not influence taxonomic practice (Schmidt-Lebuhn, 2012); nonetheless, classification decisions may have political implications, particularly in conservation (Stuessy and Hörandl, 2014). When there is a choice, and good arguments can be made for recognizing such lineages as families, the answer seems clear: recognize the family.

In the past, genera *Hydnora* and *Prosopanche* have been relatively unknown to the botanical community as their occurrence is very local and rare. However, more recently their visibility has increased as new species are discovered and described (Bolin et al., 2011; Machado and Queiroz, 2012; Martel et al., 2018; Funez et al., 2019). If Lactoridaceae are recognized, as argued above, this in turn also supports recognition of Hydnoraceae (and Asaraceae) as distinct from the more narrowly defined Aristolochiaceae. This is supported by their rather bizarre morphology that is unique among angiosperms, and is consistent with the classification of other highly modified heterotrophic plants as families, such as Rafflesiaceae. The latter was accepted as a segregate family in Malpighiales by APG et al. (2016), in contrast to its inclusion in Euphorbiaceae s.l. by APG (2009), based on the same survey by Christenhusz et al. (2015), where a majority of respondents found it "*. . .difficult to conceive an expanded Euphorbiaceae that includes a taxon as divergent*." Moreover, Hydnoraceae were not classified in Piperales until the study by Nickrent et al. (2002), which was then accepted by APG (2003). Prior to that, the family had generally been placed near Rafflesiaceae (e.g., Cronquist, 1988, who classified it in Rafflesiales, although recognizing Hydnoraceae as clearly distinctive).

Given the abovementioned arguments, we believe that the classification of perianth-bearing Piperales should therefore be reconsidered to recognize the four monophyletic families Aristolochiaceae, Asaraceae, Hydnoraceae, and Lactoridaceae.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: TreeBase (http://purl.org/phylo/treebase/phylows/study/TB2:S27866).

AUTHOR CONTRIBUTIONS

SW: conception of the study. MJ, M-SS, IM, SG, and SW: data generation. MJ: analyses and visualization of results. MJ and SW: writing of the first draft. All authors reviewed and edited the draft and agreed to the published version of the manuscript.

FUNDING

We thank the German Academic Exchange Service (DAAD) for funding exchange between Mexico (PPP Mexico) and Canada (PPP Canada), the academic exchange office of TU Dresden and the Leonardo office Dresden, as well as Erasmus+ KA107 action for travel funds and mobility organization. We thank the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program (FP7/2007-2013) under REA grant agreement no. 301257 and the NSERC Discovery program.

ACKNOWLEDGMENTS

We would like to thank Tod Stuessy and David Mabberley for their feedback on an earlier version of the manuscript, and in addition Tod Stuessy for providing a picture of *Lactoris* and Christoph Neinhuis for providing a picture of *Saruma*. We would also like to thank Jakob Wegener for help with laboratory work. We would further like to thank Yuannian Jiao and Julia Naumann for providing yet unpublished data. Last but not least, we thank the Botanical Garden of TU Dresden for cultivation of plant material and the continued support by Christoph Neinhuis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.642598/full#supplementary-material

REFERENCES

Agardh, C. (1821). *Aphorismi Botanici 7*. Lund: Berling, 87–102.

Ahmad, N., Fazal, H., Abbasi, B. H., Farooq, S., Ali, M., and Khan, M. A. (2012). Biological role of *Piper nigrum* L.(Black pepper): a review. *Asian Pacific J. Trop. Biomed.* 2, S1945–S1953.

Allio, R., Nabholz, B., Wanke, S., Chomicki, G., Pérez-Escobar, O. A., Cotton, A. M., et al. (2021). Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *Nat. Commun.* 12, 1–15.

Angiosperm Phylogeny Group (1998). An ordinal classification for the families of flowering plants. *Ann. Missouri Bot. Garden* 85, 531–553. doi: 10.2307/2992015

Angiosperm Phylogeny Group (2003). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot. J. Linnean Soc.* 141, 399–436. doi: 10.1046/j.1095-8339.2003.t01-1-00158.x

Angiosperm Phylogeny Group (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linnean Soc.* 161, 105–121. doi: 10.1111/j.1095-8339.2009.00996.x

Angiosperm Phylogeny Group, Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., et al. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linnean Soc.* 181, 1–20. doi: 10.1111/boj.12385

Backlund, A., and Bremer, K. (1998). To be or not to be – principles of classification and monotypic plant families. *TAXON* 47, 391–400. doi: 10.2307/1223768

Geneious. (2020). *Geneious 11.1.5*. Available online at: https://www.geneious.com/ (accessed November 14. 2020).

Bolin, J. F., Lupton, D., and Musselman, L. J. (2018). *Hydnora arabica* (Aristolochiaceae), a new species from the Arabian Peninsula and a key to *Hydnora*. *Phytotaxa* 338:99. doi: 10.11646/phytotaxa.338.1.8

Bolin, J. F., Maass, E., and Musselman, L. J. (2009). Pollination biology of *Hydnora africana* Thunb.(Hydnoraceae) in Namibia: brood-site mimicry with insect imprisonment. *Int. J. Plant Sci.* 170, 157–163. doi: 10.1086/593047

Bolin, J. F., Maass, E., and Musselman, L. J. (2011). A new species of *Hydnora* (Hydnoraceae) from Southern Africa. *Syst. Bot.* 36, 255–260. doi: 10.1600/036364411x569453

Christenhusz, M. J. M., Vorontsova, M. S., Fay, M. F., and Chase, M. W. (2015). Results from an online survey of family delimitation in angiosperms and ferns: recommendations to the Angiosperm Phylogeny Group for thorny problems in plant classification. *Bot. J. Linnean Soc.* 178, 501–528. doi: 10.1111/boj.12285

Cronquist, A. (1988). *The Evolution and Classification of Flowering Plants*, Second Edn. New York, NY: The New York Botanical Z Garden, Bronx, 503–517.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772–772. doi: 10.1038/nmeth.2109

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Duarte, J. M., Wall, P. K., Edger, P. P., Landherr, L. L., Ma, H., Pires, P. K., et al. (2010). Identification of shared single copy nuclear genes in *Arabidopsis, Populus, Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* 10:61. doi: 10.1186/1471-2148-10-61

Engler, A. (1887). Über die Familie der Lactoridaceae. *Bot. Jahrb. Syst.* 8, 53–56.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410. doi: 10.1093/sysbio/27.4.401

Agilent (2020). *Fragment Analyzer Systems, Agilent*. https://www.agilent.com/en/product/automated-electrophoresis/fragment-analyzer-systems (accessed November 14. 2020).

Frenzke, L., Goetghebeur, P., Neinhuis, C., Samain, M.-S., and Wanke, S. (2016). Evolution of epiphytism and fruit traits act unevenly on the diversification of the species-rich genus *Peperomia* (Piperaceae). *Front. Plant Sci.* 7:1145.

Frenzke, L., Scheiris, E., Pino, G., Symmank, L., Goetghebeur, P., Neinhuis, C., et al. (2015). A revised infrageneric classification of the genus *Peperomia* (Piperaceae). *Taxon* 64, 424–444. doi: 10.12705/643.4

Funez, L. A., Ribeiro-Nardes, W., Kossmann, T., Peroni, N., and Drechsler-Santos, E. R. (2019). *Prosopanche demogorgoni*: a new species of *Prosopanche* (Aristolochiaceae: Hydnoroideae) from southern Brazil. *Phytotaxa* 422, 93–100. doi: 10.11646/phytotaxa.422.1.6

Gamerro, J. C., and Barreda, V. (2008). New fossil record of Lactoridaceae in southern South America: a palaeobiogeographical approach. *Bot. J. Linnean Soc.* 158, 41–50. doi: 10.1111/j.1095-8339.2008.00860.x

Giseke, P. (1792). *Praelectiones in Ordines Naturales Plantanum*. Hamburgi : impensis Benj. Gottl. Hoffmanni.

Graham, S. W., and Olmstead, R. G. (2000). Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am. J. Bot.* 87, 1712–1730. doi: 10.2307/2656749

Hamid, R. A., Zakaria, N., and Zuraini, A. (2007). Anti-ulcer activity of aqueous ethanol extract of *Peperomia pellucida* in Sprague Dawley rats. *Planta Med.* 73:455.

Hendy, M. D., and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38, 297–309. doi: 10.2307/2992396

Hennig, W. (1966). *Phylogenetic Systematics*. Urbana: Univ. Illinois Press.

Horner, H. T., Samain, M.-S., Wagner, S. T., and Wanke, S. (2015). Towards uncovering evolution of lineage-specific calcium oxalate crystal patterns in Piperales. *Botany* 93, 159–169. doi: 10.1139/cjb-2014-0191

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Isnard, S., Prosperi, J., Wanke, S., Wagner, S. T., Samain, M.-S., Trueba, S., et al. (2012). Growth form evolution in Piperales and its relevance for understanding angiosperm diversification: an integrative approach combining plant architecture, anatomy, and biomechanics. *Int. J. Plant Sci.* 173, 610–639. doi: 10.1086/665821

Jaramillo, M. A., Manos, P. S., and Zimmer, E. A. (2004). Phylogenetic relationships of the perianthless piperales: reconstructing the evolution of floral development. *Int. J. Plant Sci.* 165, 403–416. doi: 10.1086/382803

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100. doi: 10.1038/nature09916

Jost, M., Naumann, J., Rocamundi, N., Cocucci, A. A., and Wanke, S. (2020). The first plastid genome of the holoparasitic Genus *Prosopanche* (Hydnoraceae). *Plants* 9:306. doi: 10.3390/plants9030306

Jussieu, A. L. (1789). Genera plantarum. *Parisiis: apud viduam Herissant et Theophilum Barrois*, 254–259.

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

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kelly, L. M., and González, F. (2003). Phylogenetic relationships in Aristolochiaceae. *Syst. Bot.* 28, 236–249.

Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* 31, 151–160. doi: 10.1007/bf02109483

Lam, V. K., Darby, H., Merckx, V. S., Lim, G., Yukawa, T., Neubig, K. M., et al. (2018). Phylogenomic inference in extremis: a case study with mycoheterotroph plastomes. *Am. J. Bot.* 105, 480–494. doi: 10.1002/ajb2.1070

Lam, V. K., Soto Gomez, M., and Graham, S. W. (2015). The highly reduced plastome of mycoheterotrophic *Sciaphila* (Triuridaceae) is colinear with its green relatives and is under strong purifying selection. *Genome Biol. Evol.* 7, 2220–2236. doi: 10.1093/gbe/evv134

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., and Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol. Biol.* 14:82. doi: 10.1186/1471-2148-14-82

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2017). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* 34, 772–773.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30, 3276–3278. doi: 10.1093/bioinformatics/btu531

Lestiboudois, G. (1826). *Botanographie Élémentaire, d'Anatomie et de Physologie Vegetale*, Vol. 453. Paris: Roret.

Mabberley, D. J. (2017). *Mabberley's Plant-Book: A Portable Dictionary of Plants, Their Classification and Uses*. Cambridge: Cambridge University Press.

Machado, R. F., and Queiroz, L. P. (2012). A new species of Prosopanche (Hydnoraceae) from northeastern Brazil. *Phytotaxa* 75, 58–64.

Macphail, M. K., Partridge, A. D., and Truswell, E. M. (1999). Fossil pollen records of the problematical primitive angiosperm family Lactoridaceae in Australia. *Plant Syst. Evol.* 214, 199–210. doi: 10.1007/bf00985739

Mamood, S. N. H., Hidayatulfathi, O., Budin, S. B., Rohi, G. A., and Zulfakar, M. H. (2017). The formulation of the essential oil of *Piper aduncum* Linnaeus (Piperales: Piperaceae) increases its efficacy as an insect repellent. *Bull. Entomol. Res.* 107:49. doi: 10.1017/s0007485316000614

Martel, C., Fernandez-Hilario, R., Tello, J. A., Arteaga, R. G., and Gerlach, G. (2018). *Prosopanche panguanensis* (Aristolochiaceae), a new species from central Peru. *Phytotaxa* 364, 241–249. doi: 10.11646/phytotaxa.364.3.3

Massoni, J., Forest, F., and Sauquet, H. (2014). Increased sampling of both genes and taxa improves resolution of phylogenetic relationships within Magnoliidae, a large and early-diverging clade of angiosperms. *Mol. Phylogenet. Evol.* 70, 84–93. doi: 10.1016/j.ympev.2013.09.010

Meng, S.-W., Chen, Z.-D., Li, D.-Z., and Liang, H.-X. (2002). Phylogeny of Saururaceae based on mitochondrial matR gene sequence data. *J. Plant Res.* 115, 0071–0076.

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES science gateway for inference of large phylogenetic trees," in *Proceedings of the 2010 Gateway Computing Environments Workshop (GCE)*, New Orleans, LA, 1–8.

Mirarab, S., and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, i44–i52.

Morales-Briones, D. F., Gehrke, B., Huang, C.-H., Liston, A., Ma, H., Marx, H. E., et al. (2020). Analysis of paralogs in target enrichment data pinpoints multiple ancient polyploidy events in *Alchemilla* sl (Rosaceae). *bioRxiv* [Preprint].

Musselman, L. J., and Visser, J. H. (1986). The strangest plant in the world. *Veld Flora* 71, 109–111.

Naumann, J., Der, J. P., Wafula, E. K., Jones, S. S., Wagner, S. T., Honaas, L. A., et al. (2016). Detecting and characterizing the highly divergent plastid genome of the nonphotosynthetic parasitic plant *Hydnora visseri* (Hydnoraceae). *Genome Biol. Evol.* 8, 345–363. doi: 10.1093/gbe/evv256

Naumann, J., Salomo, K., Der, J. P., Wafula, E. K., Bolin, J. F., Maass, E., et al. (2013). Single-copy nuclear genes place haustorial hydnoraceae within piperales and reveal a cretaceous origin of multiple parasitic angiosperm lineages. *PLoS One* 8:e79204. doi: 10.1371/journal.pone.0079204

NCBI (2020). *NCBI Nucleotide Data Base*. Available online at: https://www.ncbi.nlm.nih.gov/nuccore/ (accessed November 14. 2020).

Neinhuis, C., Wanke, S., Hilu, K. W., Müller, K., and Borsch, T. (2005). Phylogeny of Aristolochiaceae based on parsimony, likelihood, and Bayesian analyses of trnL-trnF sequences. *Plant Syst. Evol.* 250, 7–26. doi: 10.1007/s00606-004-0217-0

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nickrent, D. L. (2020). Parasitic angiosperms: how often and how many? *Taxon* 69, 5–27. doi: 10.1002/tax.12195

Nickrent, D. L., Blarer, A., Qiu, Y.-L., Soltis, D. E., Soltis, P. S., and Zanis, M. (2002). Molecular data place Hydnoraceae with Aristolochiaceae. *Am. J. Bot.* 89, 1809–1817. doi: 10.3732/ajb.89.11.1809

Nickrent, D. L., Blarer, A., Qiu, Y.-L., Vidal-Russell, R., and Anderson, F. E. (2004). Phylogenetic inference in Rafflesiales: the influence of rate heterogeneity and horizontal gene transfer. *BMC Evol. Biol.* 4, 1–17.

Nickrent, D. L., Malécot, V., Vidal-Russell, R., and Der, J. P. (2010). A revised classification of Santalales. *Taxon* 59, 538–558. doi: 10.1002/tax.592019

Nyffeler, R., and Eggli, U. (eds). (2020). "Introduction to the classification of monocotyledons," in *Monocotyledons*, Vol. 2, Berlin: Springer, 1–6. doi: 10.1007/978-3-662-56486-8_113

Oelschlägel, B., von Tschirnhaus, M., Nuss, M., Nikolić, T., Wanke, S., Dötterl, S., et al. (2016). Spatio-temporal patterns in pollination of deceptive *Aristolochia rotunda* L.(Aristolochiaceae). *Plant Biol.* 18, 928–937. doi: 10.1111/plb.12503

Oelschlägel, B., Gorb, S., Wanke, S., and Neinhuis, C. (2009). Structure and biomechanics of trapping flower trichomes and their role in the pollination biology of *Aristolochia* plants (Aristolochiaceae). *New Phytol.* 184, 988–1002. doi: 10.1111/j.1469-8137.2009.03013.x

Oelschlägel, B., Nuss, M., von Tschirnhaus, M., Pätzold, C., Neinhuis, C., Dötterl, S., et al. (2015). The betrayed thief–the extraordinary strategy of *Aristolochia rotunda* to deceive its pollinators. *New Phytol.* 206, 342–351. doi: 10.1111/nph.13210

Oelschlägel, B., Wagner, S., Salomo, K., Pradeep, N. S., Yao, T. L., Isnard, S., et al. (2011). Implications from molecular phylogenetic data for systematics, biogeography and growth form evolution of *Thottea* (Aristolochiaceae). The Gardens'. *Bull. Singapore* 63, 259–275.

Pabón-Mora, N., Madrigal, Y., Alzate, J. F., Ambrose, B. A., Ferrándiz, C., Wanke, S., et al. (2020). Evolution of Class II TCP genes in perianth bearing Piperales and their contribution to the bilateral calyx in *Aristolochia*. *New Phytol.* 228, 752–769. doi: 10.1111/nph.16719

Pabón-Mora, N., Suárez-Baron, H., Ambrose, B. A., and González, F. (2015). Flower development and perianth identity candidate genes in the basal angiosperm *Aristolochia fimbriata* (Piperales: Aristolochiaceae). *Front. Plant Sci.* 6:1095.

Qiagen (2020). *CLC Genomics Workbench*. Maryland: Qiagen.

Qiu, Y.-L., Lee, J., Bernasconi-Quadroni, F., Soltis, D. E., Soltis, P. S., Zanis, M., et al. (2000). Phylogeny of basal angiosperms: analyses of five genes from three genomes. *Int. J. Plant Sci.* 161, S3–S27.

Quijano-Abril, M. A., Callejas-Posada, R., and Miranda-Esquivel, D. R. (2006). Areas of endemism and distribution patterns for Neotropical *Piper* species (Piperaceae). *J. Biogeogr.* 33, 1266–1278. doi: 10.1111/j.1365-2699.2006.01501.x

Raju, M., and Ramesh, B. (2012). Phytochemical investigation and pharmacological activity in the roots of *Thottea siliquosa* Lam. *Asian J. Biol. Life Sci.* 1, 72–75.

Ricci, M. (2001). Evaluation of conservation status of *Lactoris fernandeziana* Philippi (Lactoridaceae) in Chile. *Biodivers. Conserv.* 10, 2129–2138. doi: 10.1023/A:1013189526734

Salomo, K., Smith, J. F., Feild, T. S., Samain, M.-S., Bond, L., Davidson, C., et al. (2017). The emergence of earliest angiosperms may be earlier than fossil evidence indicates. *Syst. Bot.* 42, 607–619. doi: 10.1600/036364417x696438

Samain, M.-S., Vrijdaghs, A., Hesse, M., Goetghebeur, P., Jiménez Rodríguez, F., Stoll, A., et al. (2010). *Verhuellia* is a segregate lineage in Piperaceae: more evidence from flower, fruit and pollen morphology, anatomy and development. *Ann. Bot.* 105, 677–688. doi: 10.1093/aob/mcq031

Sati, H., Sati, B., Saklani, S., Bhatt, P. C., and Mishra, A. P. (2011). Phytochemical and pharmacological potential of *Aristolochia indica*: a review. *Res. J. Pharm. Biol. Chem. Sci.* 2, 647–654.

Sayyari, E., and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* 33, 1654–1668. doi: 10.1093/molbev/msw079

Schmidt-Lebuhn, A. N. (2012). Fallacies and false premises—A critical assessment of the arguments for the recognition of paraphyletic taxa in botany. *Cladistics* 28, 174–187. doi: 10.1111/j.1096-0031.2011.00367.x

Seymour, R. S., Maass, E., and Bolin, J. F. (2009). Floral thermogenesis of three species of *Hydnora* (Hydnoraceae) in Africa. *Ann. Bot.* 104, 823–832. doi: 10.1093/aob/mcp168

Shah, S. K., Garg, G., Jhade, D., and Patel, N. (2016). *Piper betle*: phytochemical, pharmacological and nutritional value in health management. *Int. J. Pharm. Sci. Rev. Res.* 38, 181–189.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508. doi: 10.1080/10635150290069913

Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1114. doi: 10.1093/oxfordjournals.molbev.a026201

Silva, W. C., de Souza Martins, J. R., de Souza, H. E. M., Heinzen, H., Cesio, M. V., Mato, M., et al. (2009). Toxicity of *Piper aduncum* L.(Piperales: Piperaceae) from the Amazon forest for the cattle tick *Rhipicephalus* (Boophilus) *microplus* (Acari: Ixodidae). *Veterinary Parasitol.* 164, 267–274. doi: 10.1016/j.vetpar.2009.06.006

Sinn, B. T., Kelly, L. M., and Freudenstein, J. V. (2015). Phylogenetic relationships in *Asarum*: Effect of data partitioning and a revised classification. *Am. J. Bot.* 102, 765–779. doi: 10.3732/ajb.1400316

Smith, A. R., Pryer, K. M., Schuettpelz, E., Korall, P., Schneider, H., and Wolf, P. G. (2006). A classification for extant ferns. *Taxon* 55, 705–731. doi: 10.2307/25065646

Soltis, D. E., Soltis, P. S., Chase, M. W., Mort, M. E., Albach, D. C., Zanis, M., et al. (2000). Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. *Bot. J. Linnean Soc.* 133, 381–461. doi: 10.1006/bojl.2000.0380

Souto, R. N. P., Harada, A. Y., Andrade, E. H. A., and Maia, J. G. S. (2012). Insecticidal activity of Piper essential oils from the Amazon against the fire ant *Solenopsis saevissima* (Smith) (Hymenoptera: Formicidae). *Neotropical Entomol.* 41, 510–517. doi: 10.1007/s13744-012-0080-6

SRA (2020). *Sequence Read Archive*. Available online at: https://www.ncbi.nlm.nih.gov/sra/ (accessed November 14. 2020).

Srivastava, S. K., and Braman, D. R. (2010). The revised generic diagnosis, specific description and synonymy of the Late Cretaceous *Rosannia manika* from Alberta, Canada: Its phytogeography and affinity with family Lactoridaceae. *Rev. Palaeobot. Palynol.* 159, 2–13. doi: 10.1016/j.revpalbo.2009.10.003

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

Stevens, P. F. (2001). *Angiosperm Phylogeny Website*. Available online at: http://www.mobot.org/mobot/research/apweb/ (accessed November 14. 2020).

Stöver, B. C., and Müller, K. F. (2010). TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinform.* 11:7.

Strimmer, K., and Rambaut, A. (2002). Inferring confidence sets of possibly misspecified gene trees. Proceedings of the Royal Society of London. *Series B Biol. Sci.* 269, 137–142. doi: 10.1098/rspb.2001.1862

Stuessy, T. F. (2009a). Paradigms in biological classification (1707–2007): Has anything really changed? *Taxon* 58, 68–76. doi: 10.1002/tax.581010

Stuessy, T. F. (2009b). *Plant Taxonomy: The Systematic Evaluation of Comparative Data*. New York, NY: Columbia University Press.

Stuessy, T. F. (2013). Schools of data analysis in systematics are converging, but differences remain with formal classification. *Taxon* 62, 876–885. doi: 10.12705/625.12

Stuessy, T. F., Crawford, D. J., Anderson, G. J., and Jensen, R. J. (1998). Systematics, biogeography and conservation of Lactoridaceae. *Perspect. Plant Ecol. Evol. Syst.* 1, 267–290. doi: 10.1078/1433-8319-00062

Stuessy, T. F., Foland, K. A., Sutter, J. F., Sanders, R. W., and Silva, M. (1984). Botanical and geological significance of potassium-argon dates from the Juan Fernandez Islands. *Science* 225, 49–51. doi: 10.1126/science.225.4657.49

Stuessy, T. F., and Hörandl, E. (2014). The importance of comprehensive phylogenetic (evolutionary) classification—A response to Schmidt-Lebuhn's commentary on paraphyletic taxa. *Cladistics* 30, 291–293. doi: 10.1111/cla.12038

Stuessy, T. F., König, C., and Sepúlveda, P. L. (2014). Paraphyly and Endemic Genera of Oceanic Islands: Implications for Conservation1. *Ann. Missouri Bot. Garden* 100, 50–78. doi: 10.3417/2012087

Stuessy, T. F., Marticorena, C., Rodriguez, R. R., Crawford, D. J., and Silva, O. M. (1992). Endemism in the vascular flora of the Juan Fernández Islands. *Aliso J. Syst. Evol. Bot.* 13, 297–307. doi: 10.5642/aliso.19921302.03

Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A., and Smith, S. A. (2020). Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. *Am. J. Bot.* 107, 790–805. doi: 10.1002/ajb2.1468

Swofford, D. L. (1998). *Phylogenetic Analysis Using Parsimony*. London: London's Global University UCL.

Thorogood, C. (2019). *Hydnora*: the strangest plant in the world? *Plants People Planet* 1, 5–7. doi: 10.1002/ppp3.9

Thunberg, C. P. (1775). *Kongliga Vetenskaps Academiens Handlingar*, Vol. 36. Stockholm: Almqvist & Wiksell, 69.

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., et al. (2018). International code of nomenclature for algae, fungi, and plants (shenzhen code) adopted by the nineteenth international botanical congress shenzhen, China, july 2017. Taunus: Koeltz Botanical Books.

Vaidya, G., Lohman, D. J., and Meier, R. (2011). SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics* 27, 171–180. doi: 10.1111/j.1096-0031.2010.00329.x

Ventenat, E. P. (1799). *Tableau du Regne Végétal, Selon la Méthode de JUSSIEU*, Vol. 1. Paris: de l'Imprimerie de J. Drisonnier.

Wagner, S. T., Isnard, S., Rowe, N. P., Samain, M.-S., Neinhuis, C., and Wanke, S. (2012). Escaping the lianoid habit: Evolution of shrub-like growth forms in *Aristolochia* subgenus Isotrema (Aristolochiaceae). *Am. J. Bot.* 99, 1609–1629. doi: 10.3732/ajb.1200244

Wanke, S., Jaramillo, M. A., Borsch, T., Samain, M.-S., Quandt, D., and Neinhuis, C. (2007a). Evolution of Piperales—matK gene and trnK intron sequence data reveal lineage specific resolution contrast. *Mol. Phylogenet. Evol.* 42, 477–497. doi: 10.1016/j.ympev.2006.07.007

Wanke, S., Samain, M.-S., Vanderschaeve, L., Mathieu, G., Goetghebeur, P., and Neinhuis, C. (2006). Phylogeny of the Genus *Peperomia* (Piperaceae) Inferred from the trnK/matK Region (cpDNA). *Plant Biol.* 8, 93–102. doi: 10.1055/s-2005-873060

Wanke, S., Vanderschaeve, L., Mathieu, G., Neinhuis, C., Goetghebeur, P., and Samain, M.-S. (2007b). From forgotten taxon to a missing link? The position of the genus *Verhuellia* (Piperaceae) revealed by molecules. *Ann. Bot.* 99, 1231–1238. doi: 10.1093/aob/mcm063

Wanke, S., Mendoza, C. G., Müller, S., Guillén, A. P., Neinhuis, C., Lemmon, A. R., et al. (2017). Recalcitrant deep and shallow nodes in *Aristolochia* (Aristolochiaceae) illuminated using anchored hybrid enrichment. *Mol. Phylogenet. Evol.* 117, 111–123. doi: 10.1016/j.ympev.2017.05.014

Wisniewski, C., Bornstein, A. J., and Wood, D. L. (2019). Eating out or dining in: insect-plant interactions among several species of Piper in the Rio Abajo forest preserve, Puerto Rico. *Selbyana* 33, 1–15.

Zavada, M. S., and Benson, J. M. (1987). First fossil evidence for the primitive angiosperm family Lactoridaceae. *Am. J. Bot.* 74, 1590–1594. doi: 10.1002/j.1537-2197.1987.tb12150.x

Zavada, M. S., and Taylor, T. N. (1986). Pollen morphology of Lactoridaceae. *Plant Syst. Evol.* 154, 31–39.

Zaveri, M., Khandhar, A., Patel, S., and Patel, A. (2010). Chemistry and pharmacology of *Piper longum* L. *Int. J. Pharm. Sci. Rev. Res.* 5, 67–76.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 03 September 2021
doi: 10.3389/fpls.2021.662715

Highly Diverse Shrub Willows (*Salix* L.) Share Highly Similar Plastomes

Natascha D. Wagner[1]*, *Martin Volf*[2] *and Elvira Hörandl*[1]

[1] Department of Systematics, Biodiversity and Evolution of Plants (With Herbarium), University of Goettingen, Göttingen, Germany, [2] Biology Centre of the Czech Academy of Sciences, Institute of Entomology, Ceske Budejovice, Czechia

Plastome phylogenomics is used in a broad range of studies where single markers do not bear enough information. Phylogenetic reconstruction in the genus *Salix* is difficult due to the lack of informative characters and reticulate evolution. Here, we use a genome skimming approach to reconstruct 41 complete plastomes of 32 Eurasian and North American *Salix* species representing different lineages, different ploidy levels, and separate geographic regions. We combined our plastomes with published data from Genbank to build a comprehensive phylogeny of 61 samples (50 species) using RAxML (Randomized Axelerated Maximum Likelihood). Additionally, haplotype networks for two observed subclades were calculated, and 72 genes were tested to be under selection. The results revealed a highly conserved structure of the observed plastomes. Within the genus, we observed a variation of 1.68%, most of which separated subg. *Salix* from the subgeneric *Chamaetia/Vetrix* clade. Our data generally confirm previous plastid phylogenies, however, within *Chamaetia/Vetrix* phylogenetic results represented neither taxonomical classifications nor geographical regions. Non-coding DNA regions were responsible for most of the observed variation within subclades and 5.6% of the analyzed genes showed signals of diversifying selection. A comparison of nuclear restriction site associated DNA (RAD) sequencing and plastome data on a subset of 10 species showed discrepancies in topology and resolution. We assume that a combination of (i) a very low mutation rate due to efficient mechanisms preventing mutagenesis, (ii) reticulate evolution, including ancient and ongoing hybridization, and (iii) homoplasy has shaped plastome evolution in willows.

Keywords: ***Chamaetia/Vetrix*** **clade, Eurasia, genome skimming, North America, phylogenomics, plastome evolution**

OPEN ACCESS

Edited by:
Susann Wicke, Humboldt University of Berlin, Germany

Reviewed by:
Roswitha Schmickl, Academy of Sciences of the Czech Republic (ASCR), Czechia
Julia Bechteler, University of Bonn, Germany
Doerte Harpke, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany
Michael Gruenstaeudl, Freie Universität Berlin, Germany

***Correspondence:**
Natascha D. Wagner
natascha.wagner@uni-goettingen.de

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *01 February 2021*
Accepted: *23 July 2021*
Published: *03 September 2021*

Citation:
Wagner ND, Volf M and Hörandl E (2021) Highly Diverse Shrub Willows (Salix L.) Share Highly Similar Plastomes. Front. Plant Sci. 12:662715. doi: 10.3389/fpls.2021.662715

INTRODUCTION

Plastid markers are frequently used in plant phylogenetics because they possess several advantages over nuclear markers (Taberlet et al., 1991; Gitzendanner et al., 2018). They are haploid but occur in high copy number, which simplifies the sequencing process. Additionally, the availability of various plastid markers with different levels of molecular evolution combined with the conserved structure of the plastome makes them a popular choice for molecular systematic studies on different levels of divergence (Shaw et al., 2005, 2007; Wicke and Schneeweiss, 2015). Plastomes are usually inherited uniparentally and only rarely show recombination between differentiated plastid genomes (Wolfe and Randle, 2004; Bock et al., 2014). In combination with nuclear markers, this makes plastomes useful for the analysis of introgression, hybridization, and polyploidy. In addition, the dispersion of maternally inherited genomes occurs at shorter geographic distances than for nuclear genomes. The consequence of a reduced gene dispersal and high genetic drift in organelle genomes

is a generally pronounced geographic structure (Besnard et al., 2011). However, despite all these advantages, single plastid markers have not been able to resolve phylogenetic relationships in some lineages due to a lack of informative sites (e.g., Percy et al., 2014). The advent of next generation sequencing techniques has enabled researchers to overcome this lack of information by analyzing complete plastomes at moderate costs, e.g., via genome skimming (Straub et al., 2012; Wicke and Schneeweiss, 2015). The number of available plastomes in databases that potentially might serve as reference for read mapping has drastically increased over the last years. Thus, plastome phylogenomics has been used in a broad range of studies, e.g., in rapidly radiating groups (Barrett et al., 2014; Straub et al., 2014) and lineages with high species diversity (Huang et al., 2014; Givnish et al., 2015; Nargar et al., 2018).

The genus *Salix* L. (Salicaceae) comprises about 400–450 species of trees and shrubs mainly occurring in the Northern Hemisphere (Fang et al., 1999; Skvortsov, 1999; Argus, 2010). Willows are ecologically and economically important, e.g., for biomass production (Smart et al., 2005; Karp et al., 2011), and they are considered as keystone plants for insect diversity (Narango et al., 2020). The reconstruction of the willow phylogeny has proven to be difficult based on traditional Sanger sequencing markers, which have failed to resolve interspecific relationships (Leskinen and Alström-Rapaport, 1999; Azuma et al., 2000; Chen et al., 2010; Savage and Cavender-Bares, 2012; Barcaccia et al., 2014; Percy et al., 2014; Lauron-Moreau et al., 2015; Wu et al., 2015). Based on morphological characters, the genus is divided into three (or five) subgenera: subgenus *Salix* s.l. (including subgenera *Salix* L., *Longifoliae* (Andersson) Argus, *Protitae* Kimura, or excluding the latter two), subgenus *Chamaetia* (Dumort) Nasarov in Kom., and subgenus *Vetrix* Dumort (Skvortsov, 1999; Argus, 2010; Lauron-Moreau et al., 2015; Wu et al., 2015). Recent studies have recommended that the latter two subgenera be merged into the *Chamaetia/Vetrix* clade (Wu et al., 2015; Wagner et al., 2018, 2020, 2021). This clade comprises about three quarters of the described species diversity in *Salix* containing more than 300 species classified in about 40 sections. Previous molecular studies based on traditional markers were able to confirm the monophyly of the genus and to separate a small, basal clade of subtropical to temperate trees (subg. *Salix* s.l.) (Leskinen and Alström-Rapaport, 1999; Azuma et al., 2000; Chen et al., 2010; Savage and Cavender-Bares, 2012; Barcaccia et al., 2014; Percy et al., 2014; Lauron-Moreau et al., 2015; Wu et al., 2015). Nevertheless, they failed to resolve the relationships among species of the diverse *Chamaetia/Vetrix* clade of shrub willows due to a lack of informative sites. Percy et al. (2014) tried to interpret the lack of variation in plastid barcoding markers with either coalescence failure and incomplete lineage sorting, or a selective, trans-specific sweep for a certain haplotype. The latter idea was supported by the observation of a non-random distribution of haplotypes and of polymorphisms within genes (Percy et al., 2014). However, the authors included only four plastid loci and focused mainly North American species. Selective sweeps were also hypothesized by Huang et al. (2014) to occur in plastomes of a few tested willow species. We aim to test if this pattern can be confirmed for complete plastome data in a more comprehensive sampling of species.

While single plastid or nuclear markers have failed to resolve relationships, recently, restriction site associated DNA (RAD) sequencing has been used to resolve relationships within the *Chamaetia/Vetrix* clade, rendering all taxonomic species as distinct monophyletic lineages (Wagner et al., 2018, 2020; He et al., 2021b). However, the data contained exclusively nuclear information. The availability of additional whole plastome data would increase our understanding of reticulate evolution within the genus. Reticulate evolution could involve several processes: ancient incomplete lineage sorting, horizontal gene transfer, and/or interspecific hybridization. In case of hybridization, including the hybrid origin of allopolyploids, the position of a species will differ between phylogenies that are based on plastid data representing the maternal lineage and nuclear data reflecting biparental inheritance. By analyzing plastomes in combination with nuclear data, it is thus possible to test hypotheses on reticulate evolution (Wicke and Schneeweiss, 2015) and to gain insight into the mode of origin for polyploids. Extant hybridization and introgression is an extensively reported and studied phenomenon in *Salix* and occurs even between distantly related species (Skvortsov, 1999; Argus, 2010; Hörandl et al., 2012; Gramlich et al., 2018). Additionally, ancient hybrid origin via allopolyploidy has been demonstrated for several European species (Wagner et al., 2020). Therefore, with the incorporation of plastid data, we may gain insights into whether frequent hybridization and chloroplast capture are leading to a spread of a few dominant plastid haplotypes as assumed among subg. *Chamaetia/Vetrix* (Percy et al., 2014; Lauron-Moreau et al., 2015). Chloroplast capture (Rieseberg and Soltis, 1991) often leads to a geographic clustering of haplotypes rather than a species-specific clustering. This is especially frequent in taxa with known hybridization and introgression (e.g., *Quercus*, Pham et al., 2017).

Recently, several single *Salix* plastomes were published (e.g., Lu et al., 2019; Wu et al., 2019; Chen, 2020) and the method of complete plastome sequencing was applied to the family Salicaceae s.l. to study phylogenetic relationships and diversification of five genera with a special focus on *Salix* and *Populus* (Huang et al., 2017; Zhang et al., 2018). This set was expanded by Li et al. (2019) to 24 species representing 18 genera. However, the authors focused on higher taxonomic levels and not on subgeneric relationships. Furthermore, few of the previous accessions covered the *Chamaetia/Vetrix* clade that contains most of the willow species. In this study, we present 41 complete plastomes of 32 *Salix* species, representing 19 out of circa 40 sections, with a specific focus on Eurasian species of the diverse *Chamaetia/Vetrix* clade to analyze plastome structure and variability. We combine the data with available *Salix* plastomes from Genbank to determine the utility of complete plastomes for phylogenetic analyses. The reconstructed relationships of the genus are used to examine if the taxonomical classification and/or biogeographical distribution are reflected by plastome data. We test whether selective sweeps could have shaped the plastome diversity of willows. Furthermore, we compare the plastome to nuclear RAD sequencing data in order to discuss ancient

and recent hybridization and introgression in our target group. Finally, we discuss possible reasons that can explain the observed level of plastome variability within the genus *Salix*.

MATERIALS AND METHODS

Plant Material

For this study, we sampled 32 species (41 accessions) representing 19 sections sensu Skvortsov (1999) (**Table 1**, **Supplementary Table 1**). Four species belonged to *Salix* subg. *Salix* s.l., (s. Skvortsov) and 28 species belonged to the shrub willow clade *Chamaetia/Vetrix*. Next to sectional representation, we covered several ploidy levels. In total, we included 21 diploid, one triploid, five tetraploid, four hexaploid, and one octoploid species. The samples were collected mainly in Central Europe, however, additional samples from Spain, United Kingdom, Northern Europe, as well as the United States were included. Species were determined after Skvortsov (1999), Argus (2010), and Hörandl et al. (2012). Leaves were dried in silica gel and herbarium voucher specimens were deposited in the herbarium of the University of Goettingen (GOET) and the University of South Bohemia. For phylogenetic analyses, we integrated 20 available plastomes from Genbank (**Supplementary Table 1**).

Genome Skimming and Reference-Based Mapping

The DNA of all samples was extracted using the Qiagen DNeasy Plant Mini Kit following instructions from the manufacturer (Valencia, CA). After quality check, the DNA of 12 samples was sent to NIG - NGS Integrative Genomics Core Unit of the University Medical Center Göttingen (UMG) (https://www.humangenetik-umg.de/en/research/nig/) for library preparation and sequencing. About 1 μg DNA of each sample was used for library preparation using the PCR FreeDNA Sample Prep Kit (Illumina) followed by the Illumina TruSeq PE Cluster Kit. The 12 samples were barcoded and multiplexed. Whole genome shotgun sequencing was performed on one lane of an Illumina HiSeq 2500 platform producing 2 × 150 bp paired end reads. With this initial test set, we wanted to assess the utility of whole genome skimming for *Salix* plastome reconstruction. The sequencing libraries for the remaining samples were generated using the NEBNext® DNA Library Prep Kit following recommendations of manufacturer. The NEBNext® Multiplex Oligos for Illumina kit was used to add indices to each sample and to enrich the libraries via PCR using P5 and indexed P7 oligos. The PCR products were then purified (AMPure XP system). Whole genome shotgun sequencing was carried out on a Novaseq 6000 platform producing 2 × 150 bp paired end reads. The quality of the resulting sequencing reads was checked with FastQC v.0.10.1 (Andrews, 2010), and the reads were assembled *de novo* for a total of 36 samples using the software Fast-Plast v.1.2.8 (McKain and Wilson, 2017) under its default settings. The minimum coverage was 0.25 of the average coverage across the respective plastome. For five samples, Fast-Plast was not able to assemble the complete plastome. It is known that fragments of the plastome were transferred to the nuclear genome in *Salix* (see Huang et al., 2014). These pseudo-copies might cause problems in a *de novo* plastome assembly approach based on deep sequencing data. To obtain plastomes for the five samples, we utilized a "mapping-to-reference" approach to receive the respective plastomes. For the reference-based assembly, we used Geneious vR11 2020.2.4 (http://www.geneious.com, Kearse et al., 2012) as described in Ripma et al. (2014). The reads were mapped to the plastome of *S. purpurea* [Genbank accession NC026722].

The annotation of plastomes was done using CPGAVAS2 (Shi et al., 2019) with default settings applying the dataset containing 2,544 reference plastomes. The results were checked and edited with Geneious R11 2020.2.4 (www.geneious.com).

Phylogenetic Analyses

For final phylogenetic analyses, the sequences of the 41 produced plastomes were combined with 20 available *Salix* plastomes from Genbank resulting in a dataset comprising 61 samples (for details, see **Supplementary Table 1**). Complete plastid genomes were aligned as a single sequence with MAFFT v3 (as implemented in Geneious R11) by applying the automatic algorithm selection with a gap open penalty of 1.53 and an offset value of 0.123. One inverted repeat (IR) copy was excluded from the alignment to avoid double weighting of identical information. Because the overall variation was low, especially within the *Chamaetia/Vetrix* clade, the effects of misaligned regions in the subsequent tree topology could be pronounced (Parks et al., 2012; Duvall et al., 2020). Therefore, the alignment was optimized using Gblocks 0.91b (Castresana, 2000) with default settings (minimum number of sequences for a conserved position set to 25, minimum number of sequences for a flank position set to 40, maximum number of contiguous non-conserved positions set to 8, and minimum length of a block set to 10). The allowed gap position was set to "none" and "all," respectively, and the results of both approaches were compared. The resulting alignments were extracted in PHYLIP format and used as input for Maximum Likelihood analysis using the general time-reversible (GTR)+Γ model of nucleotide substitution implemented in RAxML (Randomized Axelerated Maximum Likelihood) v.8.2.4 (Stamatakis, 2014). We conducted for each ML analysis a rapid bootstrapping (BS) analysis with 100 replicates. Resulting trees were obtained in FigTree v1.4.3 (Rambaut, 2014).

Haplotype Networks

Next to phylogenetic tree reconstruction, we used the plastome data to calculate haplotype networks with TCS v1.21 (Clement et al., 2000). Due to the large genetic distance between the subclades, we conducted two separate analyses without outgroup: one for the closely related *Chamaetia/Vetrix* clade and one for subgenus *Salix* s.l. We used only coding regions to avoid homoplasy in the data set. Gaps were treated as missing data.

Statistical Tests

A geographic clustering rather than a taxonomic clustering is frequently observed in plastid-based studies (Gitzendanner et al., 2018). To test for this trend in our data, we correlated the genetic distance with the geographic distance of the included samples. We derived the genetic distance matrix from the branch lengths in the observed RAxML tree of the complete plastome

TABLE 1 | Plant material including taxonomic classification and origin.

Species	Subgenus	Section	Ploidy	Sample ID	Origin
Salix appendiculata	*Chamaetia/Vetrix*	*Vetrix* subs. *Vulpinae*	2x	NW17.021	Austria
Salix acutifolia	*Chamaetia/Vetrix*	*Daphnella*	2x	ACU 1	Czech Republic
Salix aurita	*Chamaetia/Vetrix*	*Vetrix* subs. *Leaves*	2x	NW17.041	Austria
				AUR4	Czech Republic
Salix bicolor	*Chamaetia/Vetrix*	*Phylicifoliae*	3x	BIC3	Austria
Salix breviserrata	*Chamaetia/Vetrix*	*Myrtosalix*	2x	EH 10508	Spain
				BRE15	Austria
Salix caesia	*Chamaetia/Vetrix*	*Helix*	4x	CAE1	Austria
Salix caprea	*Chamaetia/Vetrix*	*Vetrix* subs. *Leaves*	2x	CP03	UK
				CAP2	Czech Republic
Salix cinerea	*Chamaetia/Vetrix*	*Vetrix* subs. *Leaves*	4x	NW17.082	Austria
				CIN1	Czech Republic
Salix daphnoides	*Chamaetia/Vetrix*	*Daphnella*	2x	DAP1	Czech Republic
Salix eleagnos	*Chamaetia/Vetrix*	*Cabae*	2x	EH 10495	Spain
Salix foetida	*Chamaetia/Vetrix*	*Villosae*	2x	FOE11	Austria
Salix glabra	*Chamaetia/Vetrix*	*Glabrella*	6x	GLA2	Austria
Salix glaucosericea	*Chamaetia/Vetrix*	*Glaucae*	8x	GSR7	Austria
Salix hastata	*Chamaetia/Vetrix*	*Hastatae*	2x	HAS3A	Austria
Salix helvetica	*Chamaetia/Vetrix*	*Villosae*	2x	1/2014	Switzerland
				HEL7	Austria
Salix herbaceae	*Chamaetia/Vetrix*	*Retusae*	2x	HER5	Austria
Salix lapponum	*Chamaetia/Vetrix*	*Villosae*	2x	LAP1	Czech Republic
Salix mielichhoferi	*Chamaetia/Vetrix*	*Nigricantes*	6x	MIE5	Austria
Salix myrsinifolia	*Chamaetia/Vetrix*	*Nigricantes*	6x	NW17.054	Austria
				MYS5	Austria
Salix myrtilloides	*Chamaetia/Vetrix*	*Myrtosalix*	2x	MYR1	Czech Republic
Salix rosmarinifolia	*Chamaetia/Vetrix*	*Incubaceae*	2x	ROS3	Czech Republic
Salix reticulata	*Chamaetia/Vetrix*	*Chamaetia*	2x	EH 10397	Italy
				RTI1	Austria
Salix retusa	*Chamaetia/Vetrix*	*Incubaceae*	6x	RET6	Austria
Salix serpillifolia	*Chamaetia/Vetrix*	*Incubaceae*	2x	SER22	Austria
Salix silesiaca	*Chamaetia/Vetrix*	*Vetrix* subs. *Vulpinae*	2x	SIL22	Czech Republic
Salix sitchensis	*Chamaetia/Vetrix*	*Sitchenses*	2x	NW18.046	California, USA
Salix viminalis	*Chamaetia/Vetrix*	*Vimen*	2x	VIM1	Czech Republic
Salix waldsteiniana	*Chamaetia/Vetrix*	*Villosae*	2x	WAL31	Austria
Salix triandra	*Salix* s.l.	*Amygdalinae*	2x	TRI4	Czech Republic
Salix alba	*Salix* s.l.	*Salix*	4x	EH 10431	Germany
				ALB8	Czech Republic
Salix fragilis	*Salix* s.l.	*Salix*	4x	FRA2	Czech Republic
Salix pentandra	*Salix* s.l.	*Pentandrae*	4x	EH 10470	Finland
				PEN3	Czech Republic

Herbarium vouchers of collected samples are deposited at the herbarium of the University of Goettingen (GOET). Detailed information is given in ***Supplementary Table 1****.*

dataset using the R package ape 5.0 (Paradis and Schliep, 2019). We calculated the geographic distance matrix based on the global positioning system (GPS) coordinates of our samples. For Genbank samples without detailed information on sampling localities, we used the distribution center of the species or the location of the institute that performed the analyses instead. We then correlated the two matrices using a Mantel test based on Pearson's product-moment correlation with 999 permutations in the R package vegan 2.5 (Oksanen et al., 2019). We performed the analysis with the full dataset and with subg. *Salix* and subg. *Chamaetia/Vetrix* clades separately. Furthermore, in the case of subg. *Chamaetia/Vetrix*, we also performed an analysis excluding its basal members (*S. arbutifolia, S. rorida, S. magnifica,* and *S. oreinoma*) that showed plastomes most divergent from the rest of the subgenus. All analyses were performed in R 3.6.1 (R Core Team, 2019).

To analyze if plastid genes are under selection, we calculated gene-wise ω (dN/dS ratios = non-synonymus vs. synonymous substitutions) with codeML implemented in paml version 4.8 (Yang, 2007). We used the model = 0 option, i.e., a single omega for the whole tree. We extracted the coding sequences (=CDS) of 72 genes (**Supplementary Table 2**) out of 78 genes in total and used the alignments and the RAxML tree of the complete sample set as input. Annotations of Genbank accessions were not complete in all cases and for statistical reasons we included only genes that were present/annotated in at least 60 of the 61 samples for this test.

Comparison to Nuclear RAD Sequencing Data of a Comparative Subset

To evaluate the phylogenetic resolution and topology of the plastome phylogeny, we compared 10 samples of our plastome data to a comparative sampling of already published RAD sequencing data. The RAD sequencing data are available at Genbank (Bioproject PRJNA433286). Previous RAD sequencing studies included two to four individuals per species and rendered species as monophyletic lineages (Wagner et al., 2018, 2020), and hence we used only one representative sample per species for the RAD sequencing analysis here. *Salix triandra* was used as an outgroup in both datasets. For the comparison, we used a subset of 10 representative plastomes of the *Chamaetia/Vetrix* clade. We used the same accessions in both datasets whenever possible. However, in two cases, due to the low amount of extracted DNA, we substituted the species in the RAD analysis with another individual of the same species. The reduced RAD sequencing set was analyzed with ipyrad v7.24 (Eaton and Overcast, 2016) using the same settings as described in Wagner et al. (2018). The minimum number of samples sharing a locus was set to 4, the maximum number of single nucleotide polymorphism(s) SNP(s) per locus was set to 20, and the maximum number of indels per locus was set to 8. With respect to the mixed ploidy of the dataset, we used a maximum of four alleles in the settings of the ipyrad pipeline (for more details see Wagner et al., 2020). Maximum Likelihood analyses for both the concatenated RAD loci as well as the plastomes were performed as described above.

RESULTS

Plastome Reconstruction

The shotgun sequencing revealed an average of 71.65 Mio raw paired reads per sample. An average of 6.86 Mio paired reads mapped to the plastome. The average coverage was 7,733 reads. The plastome lengths varied between 155,414 bp (*S. mielichhoferi*) and 160,386 bp (*S. myrtilloides*). Length variation was due to one large insertion at the margin of the inverted repeat (IRb) that was observed in 21 samples, two large indels (>200 bp) in *S. triandra* and species of subg. *Salix,* several smaller indels (2–80 bp), and repetitive motifs (SSRs, tandem repeats). All obtained plastomes showed the typical tetrapartite structure of two IRs separating the small single-copy (SSC) region from the large single-copy (LSC) region. They contained 78 protein coding genes, 30 tRNAs, and 3 rRNAs. The order of genes was identical in all newly assembled plastomes. The annotated plastomes were uploaded to Genbank, their respective accession numbers (MW435413–MW435453) are provided in **Supplementary Table 1**.

Phylogenetic Reconstructions

The plastome sequences presented here were aligned together with 20 available *Salix* plastomes from Genbank. The initial alignment of the complete plastomes of 61 samples had a length of 141,081 bp and after trimming of one IR and alignment optimization with Gblocks retained a length of 129,052 bp. The concatenated alignment of coding regions (CDS) had a length of 68,211 bp. The length of the edited alignment for the *Chamaetia/Vetrix* clade was 128,608 bp and 68,009 bp for the extracted coding regions. The lengths were 128,403 bp and 68,311 bp for subgenus *Salix,* respectively. The variation observed in the complete alignment was 1.68 and 0.72% in coding regions. Within the *Chamaetia/Vetrix* clade, 0.74% of sites were variable and we observed 0.41% variability in the alignment of extracted coding regions. Within subgenus *Salix,* we observed 0.64% variability and 0.35% of variable sites for CDS, respectively. A statistical comparison of the different alignment editing approaches is provided in **Supplementary Table 3**.

Relationships of Genus *Salix* Based on Plastome Data

The observed phylogenetic tree based on complete plastomes showed a clear separation of the subgenus *Salix* s.l. (tree willows) (BS 100) and the *Chamaetia/Vetrix* clade (shrub willows) plus *S. triandra* (BS 100) (**Figure 1**). Both accessions of *S. triandra* (BS 99) were found to be sister to the well-supported *Chamaetia/Vetrix* clade (BS 98). Within the *Chamaetia/Vetrix* clade, the observed resolution was low, indicated by short branches and no or low BS support for most branches. *Salix arbutifolia* was in sister position to the remaining samples of *Chamaetia/Vetrix* (BS 100), followed by the Asian species *S. rorida*, and a clade comprising *S. magnifica* and *S. oreinoma* (BS 71). The remaining samples formed a well-supported clade (BS 100) with *Salix retusa* (RET6) at an early diverging position. The two accessions of *S. myrsinifolia* grouped together with high support (BS 100), while all other species with more than one sample appeared polyphyletic. Additionally, the sectional classification was not reflected by the phylogeny. No geographical pattern was observed, e.g., *Salix sitchensis* from California was shown to be closely related to European *S. myrsinifolia* and *S. caesia* (BS 92). Asian *S. gracilistyla* appeared in close relationship to *S. waldsteiniana* and *S. breviserrata,* both occurring in the European Alps. However, resolution within this clade was extremely low. Within the subg. /*Salix*/ clade, *S. babylonica* was in sister relationship to a subclade (BS 100) containing the European tree species (*S. alba, S. pentandra*, and *S. fragilis)*, and *S. paraplesia* from China. The North American willow *S. interior* was shown to be situated on a long branch and in sister position to the remaining samples of subg. *Salix*.

Haplotype Networks

We calculated haplotype networks of the two subclades based on coding regions (CDS). For subg. *Salix,* the haplotype

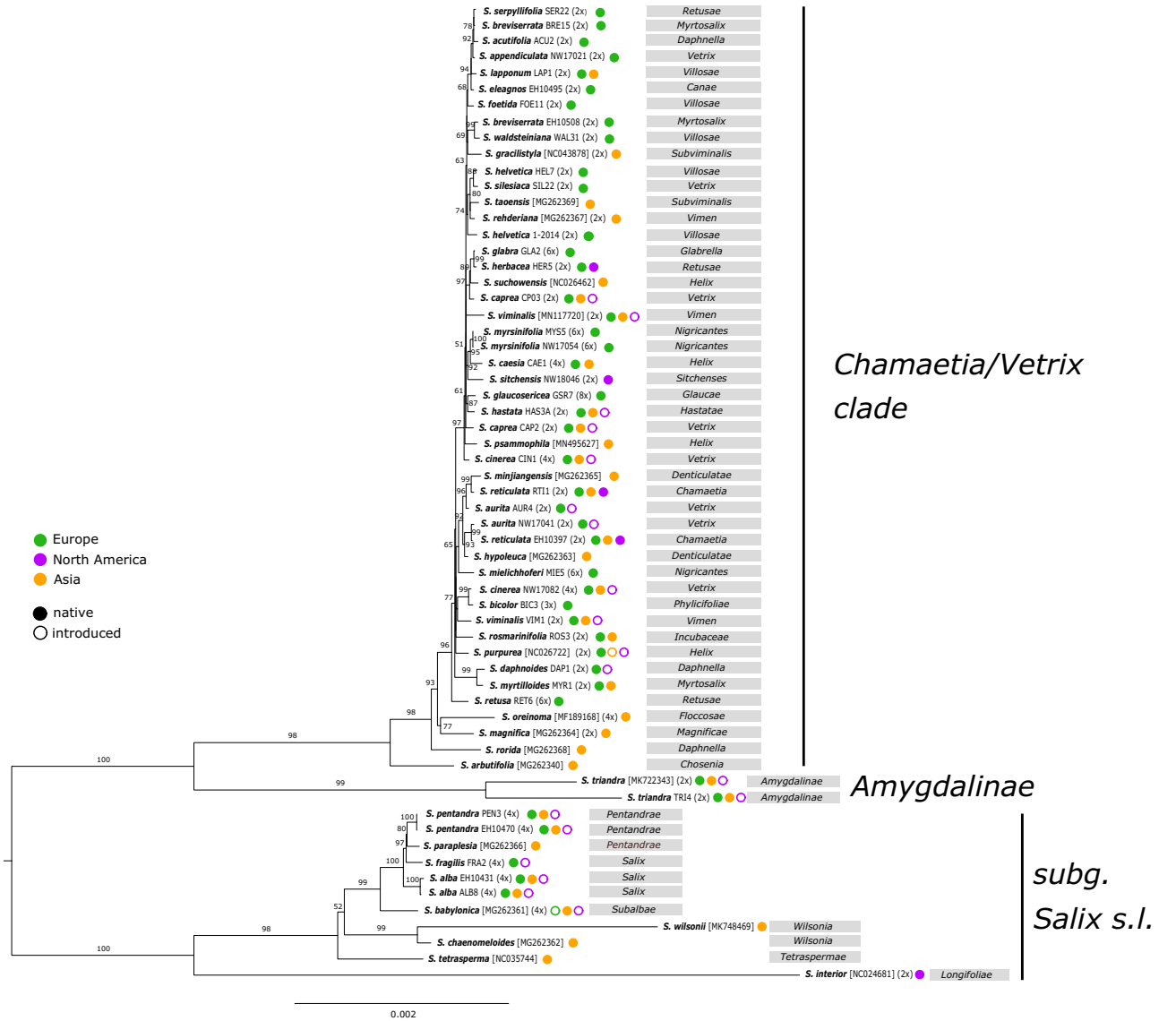


FIGURE 1 | RAxML phylogeny based on complete plastomes of 61 samples representing 50 *Salix* species. Sample IDs or Genbank accession numbers for Genbank samples as well as ploidy levels are indicated behind species names. Bootstrap support values are given above branches. Geographic distribution of species is indicated behind sample names according to color coding. The taxonomical affiliation into sections is given in gray boxes. The two main clades are illustrated according to their subgeneric classification.

network displayed the same clades as the RAxML tree (**Supplementary Figure 1**). The close relationships of tree willows collected in Europe were reflected by a central empty haplotype that was not occupied by any included species. Both *S. alba* accessions were only one mutational step apart from each other, both accessions of *S. pentandra* shared an identical haplotype. *Salix paraplesia* showed close relationships to *S. alba, S. fragilis,* and *S. pentandra*. The remaining Asian species were up to 92 mutational steps distinct from the central haplotype connecting the European tree willows plus *S. paraplesia*.

The haplotype network for the *Chamaetia/Vetrix* clade based on coding regions revealed that both accessions of *S. myrsinifolia* share the same haplotype. All other included samples showed a unique haplotype. Samples of the same species but from different localities did not group together, e.g., *S. caprea* from the United Kingdom was three mutational steps apart from *S. caprea* collected in the Czech Republic. Overall, the network showed three main groups that were connected via central haplotypes that were not occupied by included samples (**Figure 2**). Neither a geographical nor a taxonomical pattern was reflected.

Statistical Tests

The Mantel tests revealed correlation of geographic and genetic distance in the case of the whole dataset ($r = 0.1705$ $p = 0.018$), for the subgenus *Salix* ($r = 0.5025$, $p = 0.001$), and for the whole

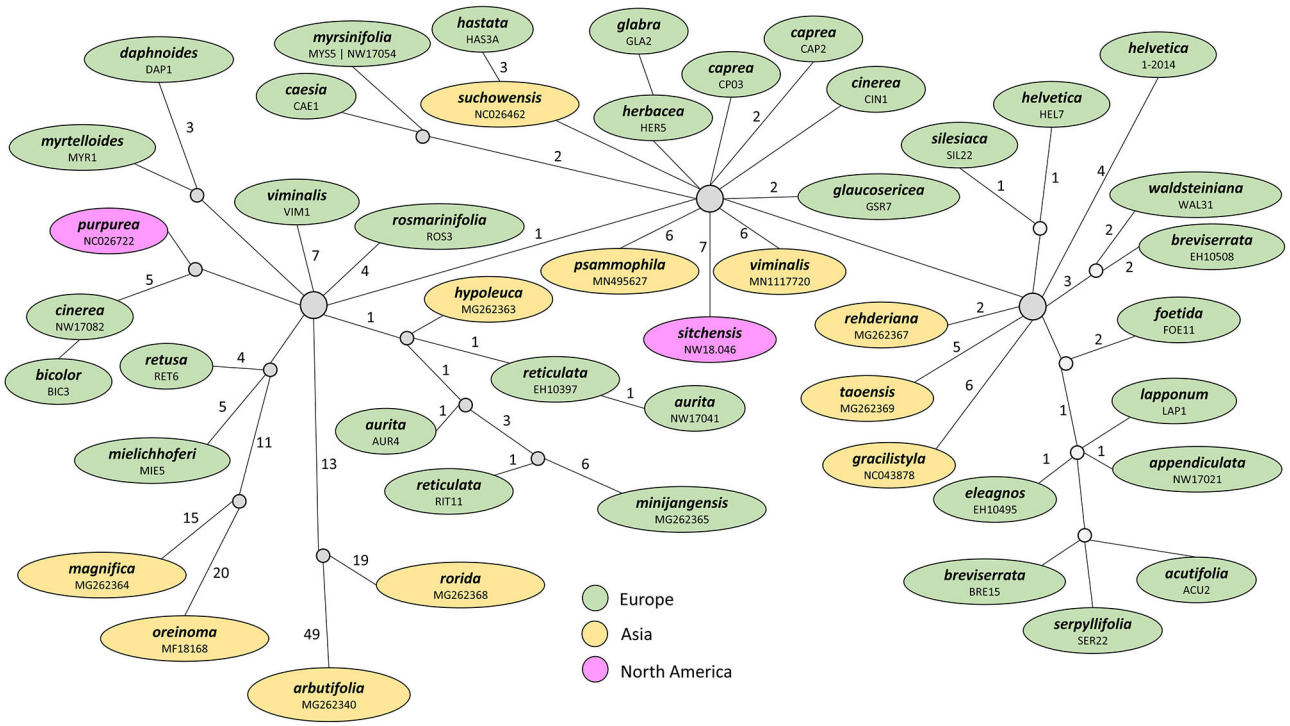


FIGURE 2 | Haplotype network of 48 samples representing 42 species of the *Chamaetia/Vetrix* clade based on CDS regions. Each circle represents a unique haplotype, the number of mutational steps between haplotypes are indicated by numbers above branches. The applied color-coding reflects the geographic origin of samples. Both accessions of *S. myrsinifolia* share an identical haplotype. The colors highlight central haplotypes and closely related groups.

subgenus *Chamaetia/Vetrix* clade ($r = 0.2739$, $p = 0.018$). When we excluded the early branching lineages of the *Chamaetia/Vetrix* clade, the significant correlation disappeared ($r = 0.0619$, $p = 0.259$).

The dN/dS ratios (non-synonymous vs. synonymous substitutions) revealed purifying selection for the majority of genes (ω values <1). Four genes showed ratios of positive, i.e., diversifying selection (*rpl*2, *rpl*16, *rps*15, and *ycf*1). A complete list of gene-wise statistics and ω values is given in **Supplementary Table 2**.

Comparison of Plastome and RAD Sequencing Data

We compared the resulting phylogeny of 10 plastomes of the *Chamaetia/Vetrix* clade to published nuclear RAD sequencing data of the same subset of species. *Salix triandra* was used as an outgroup. The resulting plastome alignment of 130 kbp showed 0.5% variable sites. The RAD sequencing alignment of 57,084 concatenated RAD sequencing loci had a length of 4,669,722 bp and showed 8.05% variable sites. The alignment contained 26.3% missing data. The observed phylogeny based on RAD sequencing data was in accordance with formerly published data based on more samples (Wagner et al., 2018, 2020, 2021) (**Figure 3**). *Salix reticulata* was in sister position to the remaining species. Species belonging to section *Vetrix* formed a well-supported monophyletic group. The North American species *S. sitchensis* was situated in sister relationship to *S. helvetica*. The plastid phylogeny of the same subset showed tetraploid *S. cinerea* in sister position to the remaining species (**Figure 3**). The members of section *Vetrix* did not form a monophylum but occurred at different positions in the tree (highlighted in **Figure 3**). *Salix sitchensis* was in sister position to hexaploid *S. myrsinifolia*. The dwarf shrub *S. breviserrata* was found to be closely related to the Swiss willow *S. helvetica*. Overall, the branches were very short, however, the bootstrap values showed moderate to good support for the observed topology.

DISCUSSION

In 2014, Percy et al. proclaimed the "spectacular failure of barcodes for willows." In contrast to single markers, we sought to analyze the variability of complete plastomes and their utility for phylogenetic analyses. We included 41 newly assembled plastomes and present here the first comprehensive study on plastome evolution on a subgeneric level within *Salix*. Compared to other angiosperms, our results reveal that the sequence variation in genus *Salix* is very low. Wu et al. (2015) observed 7.8% variable sites in their combined four-cp-marker set. However, the authors included next to *Salix* also samples of *Populus* and *Dovyales* as outgroups, which were responsible for most of the observed variation. To compare the results in detail, we extracted the four specific loci (*mat*K, *rbc*L, *atp*B-*rbc*L, and *trn*D-*trn*T) of our dataset and revealed 3.2% variable sites for the combined loci and complete sampling, but only 0.4% for the *Chamaetia/Vetrix* clade and 2.4% for subg. *Salix*, respectively (see **Supplementary Table 4**).

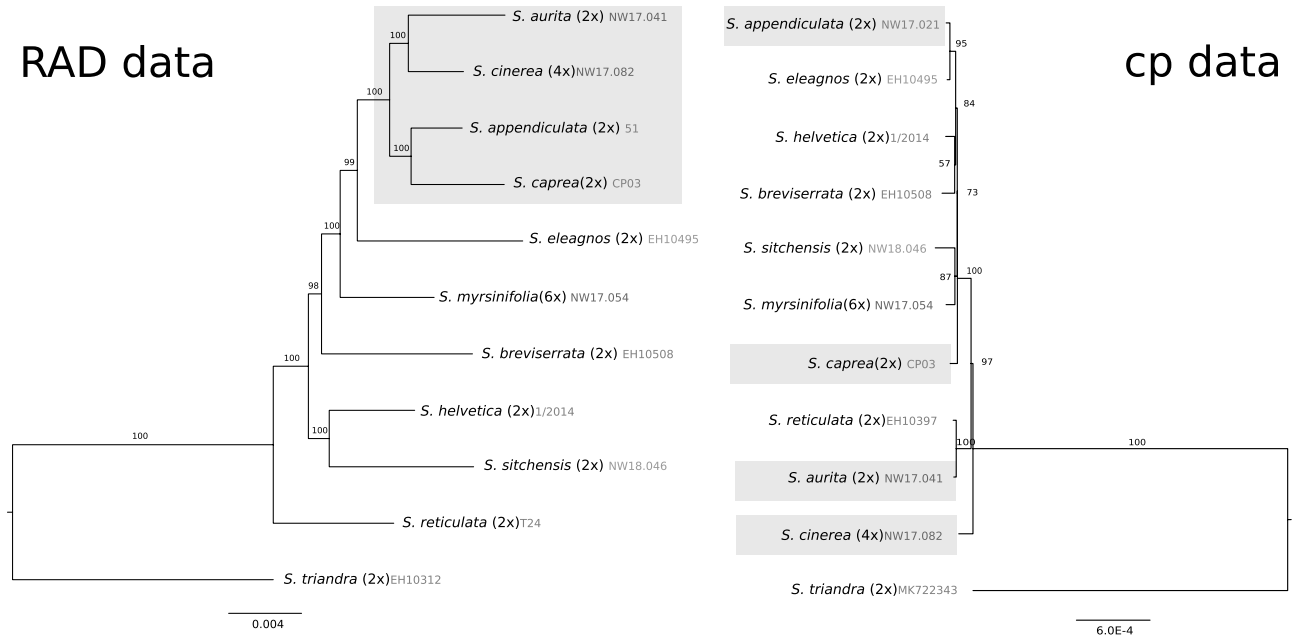


FIGURE 3 | Comparison of Maximum Likelihood phylogenies of a subset of 10 species of the *Chamaetia/Vetrix* clade with *S. triandra* as out-group based on 57,084 RAD sequencing loci **(left)** and whole plastome data **(right)**. RAD sequencing data from Wagner et al. (2018, 2020). Bootstrap support values are given above branches. The ploidy level and accession numbers are indicated behind species names, the members of section *Vetrix* are highlighted in gray.

Low Mutation Rates and Effective Repair Mechanisms in Coding Regions

Our alignment of the complete plastomes shows only 1.64% variable characters. Our alignment of all extracted coding regions shows only a variation of 0.72%. Genic regions of plastomes evolve with only about a third or half the rate of the nuclear genome (Wolfe et al., 1987). It is still unknown, why plant organellar genes have lower mutation rates than nuclear genes, but possible explanations include differences in replication enzymes, replication fidelities, mismatch repair, and low rates of genetic exchange (Gaut et al., 2011). Our results show a high degree of purifying selection in the protein coding genes of the plastome (**Supplementary Table 2**). The elimination of deleterious mutations might also affect linked sites and thus decrease the overall genetic variability (Charlesworth et al., 1993). Because the complete plastome can be treated as one single haplotype, this might be an explanation of the observed low variability. Next, recombination is a driver of purifying selection and can also happen between and among organelles (Bock et al., 2014). However, we assume that gene conversion might be a strong mechanism acting toward purifying selection in our dataset (Wolfe and Randle, 2004). Gene conversion is known from non-recombining systems, e.g., mitochondrial genomes (Mower et al., 2010) and nuclear genomes of ancient asexual animals (Flot et al., 2013). However, the elimination of deleterious mutations by gene conversion has also been proposed in plastid genomes (Khakhlova and Bock, 2006).

An overall low genetic divergence also occurs in the nuclear genomes of willows. Single barcoding regions like internal transcribed spacer (ITS) or single genes like *rbc*L and *mat*K have failed to resolve interspecific relationships (Leskinen and Alström-Rapaport, 1999; Lauron-Moreau et al., 2015). Instead, thousands of nuclear RAD sequencing loci were required to resolve species-level relationships in the *Chamaetia/Vetrix* clade (Wagner et al., 2018, 2020, 2021; He et al., 2021b). Based on our results, as well as on previous studies, we infer generally low mutation rates in willow genomes, considering the relatively high age of the genus (up to 43.8 Ma; Wu et al., 2015). Efficient regulation of intracellular oxidative stress resulting from photosynthesis and respiration might avoid DNA damage and reduce frequencies of non-homologous DNA repair processes, which is generally a major source for mutagenesis (Friedberg and Meira, 2006). Willows are rich in antioxidants, especially in phenolics and other typical chemical compounds known for the regulation of redox homeostasis (Hörandl et al., 2012; Jia et al., 2020; Piatczak et al., 2020). Their hypothetical role in the observed low mutation rates would need to be tested. However, it is remarkable that a low mutation rate (c. one-sixth of *Arabidopsis*) has also been observed in nuclear, plastid, and mitochondrial genomes of poplar (Tuskan et al., 2006), the sister genus of *Salix,* which is similarly rich in phenolics, such as salicylates, tannins, or flavonoids (Palo, 1984).

Variable but Not Informative: Rapidly Evolving Non-coding Regions

In our dataset, we observed some length variation based on insertions/deletions resulting from sequence duplications in non-coding regions of the plastome, which is in the range of similar

studies (Huang et al., 2017; Zhang et al., 2018; Li et al., 2019). Most of the observed variable characters occurred in rapidly evolving, non-coding parts of the plastome, as SSRs and other repetitive regions (Zhang et al., 2018; Li et al., 2019). Next to that, the haplotype network of the *Chamaetia/Vetrix* clade revealed mainly synonymous and non-directional mutations in coding regions (**Figure 2**). Homoplasy might be introduced by plastid haplotype polymorphism within and among individuals, resulting in paralogous copies (Wolfe and Randle, 2004). Further, the non-directional signal of mutations might lead to conflicting signals in the phylogeny (Parks et al., 2012; Duvall et al., 2020). Because the overall variability is very low, this effect might be even stronger within shrub willows. However, both the effects, low variation and non-directional signal, lead in combination to a non-resolved tree, especially in the *Chamaetia/Vetrix* clade. Interestingly, the effects in subgenus *Salix* seem to be less significant. Despite similar levels of variability, the topology of the subclade is much better resolved. This is in accordance with previous phylogenetic studies (Percy et al., 2014; Lauron-Moreau et al., 2015; Wu et al., 2015). Gene transfer from the plastome to the nucleus might give some additional explanation to the observed low variability (Bock and Timmis, 2008). For *Populus trichocarpa,* the transfer of the whole plastome to the nuclear genome was reported, while hints of transfer of single loci were also found in some *Salix* species (Huang et al., 2014). However, due to the lack of suitable genomes, we did not test for any transfer of plastid genes or larger portions of the plastome to the nuclear genome in our dataset. Nevertheless, the *de novo* assembly problems for five samples may have occurred due to the transfer of large portions of the plastome to the nucleus.

Molecular Dating Opposed Hypotheses of Rapid Radiation or Rapid Range Expansion From Refugia

Another explanation for the observed low plastome variation, especially within the *Chamaetia/Vetrix* clade, might be a large radiation or a rapid range expansion from refugia after the last glacial maximum (Percy et al., 2014; Lauron-Moreau et al., 2015). In this scenario, recently evolved species would share identical haplotypes. Our data on shrub willows revealed a low amount of variation, but almost no identical haplotypes were observed. Additionally, lineage diversification clearly predates the Pleistocene glaciations; the age of genus *Salix* was estimated as 43.8 Ma, and that of the *Chamaetia/Vetrix* clade as 23 Ma, respectively (Wu et al., 2015). Next to the diversification time, the distinctiveness of the species based on morphology and nuclear phylogenies also oppose a postglacial rapid radiation as an explanation for low plastome variability (Wagner et al., 2018, 2020; He et al., 2021b). However, an older radiation followed by fragmentation and genetic drift cannot be ruled out completely.

Differences Between Tree Willows (subg. *Salix*) and Shrub Willows (subg. *Chamaetia/Vetrix*)

Our comprehensive plastome phylogeny confirmed the differentiation into two distinct subgeneric clades (Wu et al., 2015; Huang et al., 2017; Zhang et al., 2018). An explanation for the split into two clades might be that species within subgenera and within sections hybridize more frequently than species between different subgenera (Hörandl, 1992). A recent study analyzed differences in sex determination systems in subg. *Salix* and the *Chamaetia/Vetrix* clade, which might be responsible for incompatibilities between the two subgeneric clades (He et al., 2021a). This would support our conclusion that the plastomes of the subgenera evolved more independently. Our data confirm the monophyly of species as well as the split of a New World and an Old World clade within subgenus *Salix* (Chen et al., 2010; Percy et al., 2014; Wu et al., 2015). The geographic pattern was further supported by the results of the Mantel test. Within the shrub willows, the somewhat isolated position of *S. arbutifolia* is in accordance with previous studies (Lauron-Moreau et al., 2015; Wu et al., 2015). The early branching Asian lineages corresponded to the Hengduan Mountain clade described in He et al. (2021b). Within the core *Chamaetia/Vetrix* clade, neither species-specific patterns nor support for previous sectional classification were found. The polyphyly of four species might be explained by homoplasy of plastid polymorphisms (see above).

Comparison of Nuclear and Plastid Data

We compared nuclear RAD sequencing data with plastome data for a subset of 10 samples. So far, few studies have performed a statistical comparison of the two reduced representation methods we used here (Pham et al., 2017; Mu et al., 2020). Our comparison clearly shows that RAD sequencing is much more efficient in resolving relationships within *Salix* than plastome data. The included members of the section *Vetrix* showed a well-supported monophyletic group in the RAD sequencing dataset, which is in accordance with previously published data (Wagner et al., 2018, 2020, 2021). However, the same species were scattered over the non-resolved tree in the plastome phylogeny (**Figure 3**). Maternal inheritance of plastomes might explain the observed incongruence of nuclear (RAD sequencing) and plastome phylogenies. These discrepancies could be further explained by chloroplast capture, and in the case of polyploid *S. cinerea*, by an allopolyploid origin (Wagner et al., 2020). In willows c. 40% of species are polyploid (Suda and Argus, 1968), which means that frequent allopolyploidy could have a major impact on phylogenetic relationships. Different ancient hybrid origins will influence the backbone of the plastome tree and thus explain discordance between nuclear and plastid phylogenies. Further, more recent hybridization or introgression events, even if infrequent, could occur between distantly related species, transferring the few and randomly appearing plastome polymorphisms to different genomic backgrounds of species.

The Lack of Any Biogeographical Pattern

In the presence of frequent hybridization as well as chloroplast capture, we would expect a biogeographic signal in the phylogeny and/or haplotype networks. Although our sampling represented mainly European species, it also included samples from several parts of Eurasia and North America. Based on our data, species of geographical proximity show distinct haplotypes while some species from separate continents share similar

plastomes. For example, *S. sitchensis* from the West Coast of the United States shared a similar plastome with *S. caesia* and *S. myrsinifolia* from Europe. Over these huge distances, extant hybridization and frequent maternal gene flow via seeds is unlikely. However, within Eurasia, distribution areas of widespread species are often overlapping (see Skvortsov, 1999), and hybridization appears possible. Interestingly, the Mantel test revealed significant correlation of geographical and genetic distance in our dataset. This can be explained by the early branching lineages from China, which show quite distinct plastomes and might influence the results. When analyzing only the core clade of shrub willows, no correlation could be observed. This is in accordance with former results of Percy et al. (2014) who found correlation of geographical and genetic distance within the overall dataset, but no correlation within a large clade of shrub willows. The close relationship of Eurasian and North American shrub willow species in plastid-based phylogenies was also reported in Lauron-Moreau et al. (2015), who observed a large clade comprising boreo-arctic and montane to alpine species of both Eurasia and North America.

Plastid Genes Under Selection

Percy et al. (2014) assumed that hybridization/ introgression alone could not explain the small number of shared haplotypes between a large number of distinct willow morphospecies. They assumed a trans-specific selective sweep as a potential reason for one dominant haplotype. The positively selected plastome would have been able to spread rapidly, probably aided by widespread species hybridizing with the local ones. Our haplotype network of the *Chamaetia/Vetrix* clade, however, does not support the predominance of one certain haplotype. Further, most of the observed variation occurred in non-coding regions, and within genes, mostly in synonymous sites. However, the scenario of a selective sweep would require a positive selection of plastid genes. Indeed, Huang et al. (2017) tested plastid coding regions and found seven genes under selection in Salicaceae. However, this is not reflected in our results. Most tested protein coding genes showed purifying selection ($dN/dS < 1$). Only four genes (5.6%) showed signals of positive selection (*rpl*2, *rpl*16, *rps*15, and *ycf*1). Interestingly, they differed from the selective genes found by Huang et al. (2017). The genes analyzed by Percy et al. (2014) (*mat*K, *rbc*L, *rpo*B, and *rpo*C1) were all under purifying selection with ω values far below one. Thus, our results based on more species strongly contradict the hypothesis of a selective sweep. In *rpl*2 and *rps*15, slight signals of positive selection were detected, but both are ribosomal genes. The signal of positive selection was strong ($dN/dS > 1$) only in the case of *ycf*1. This large open reading frame was for a long time enigmatic and its function unknown. The gene *ycf*1 has been predicted to have the highest nucleotide diversity (π) at the species level within angiosperm plastid genomes (Dong et al., 2012, 2015). More recently, it was shown that *ycf*1 encodes for Tic214, a vital component of the *Arabidopsis* translocon on the inner chloroplast (TIC) membrane complex that is essential for plant viability (Kikuchi et al., 2013). However, in comparison to other plant genera, *ycf*1 is relatively conserved and showed only 1.5% variability on the genus level within *Salix*. The lack of a predominant haplotype as well as the low number of genes under selection argue against the hypotheses of a selective sweep in willows.

CONCLUSIONS

The observed plastome variation in willows is much lower than in other angiosperm lineages. Thus, even complete plastome data are unsuitable for phylogenetic reconstruction, DNA barcoding, and analyses of biogeographical history in shrub willows. Usual explanations for plastome evolution patterns do not fit our data. Instead, the willow plastomes seem to have been shaped by extremely low mutation rates due to efficient mechanisms preventing mutagenesis, and further, by reticulate evolution and non-specific, rather random polymorphisms resulting in homoplasy. Consequently, the observed plastomes are neither species-specific nor reflect geographical patterns. Our results provide a caveat on relying solely on plastid phylogenies, a common practice in plant systematics. Our study demonstrates the importance of examining the evolution of plastid genomes thoroughly before applying them to questions of plant systematics, especially in cases of widespread, hybridizing taxa with low evolutionary rates.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/genbank/, MW435413 - MW435453.

AUTHOR CONTRIBUTIONS

NW planned and designed research. NW, MV, and EH conducted fieldwork. NW and MV performed experiments and analyzed data. NW wrote the manuscript. MV and EH contributed to the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was financially supported by the German research foundation Deutsche Forschungsgemeinschaft, DFG (Ho 5462/7-1 to EH and priority program Taxon-Omics SPP1991, project Wa3684/2-1 to NW). MV acknowledges funding by Czech Academy of Sciences, Programme for Research and Mobility Support of Starting Researchers (MSM200962004), and Grant Agency of the Czech Republic (20-10543Y).

ACKNOWLEDGMENTS

We thank Bruce Baldwin, Susanne Gramlich, Andrea Danler, Katrin Scheufler, Claudia Pätzold, Carlo L. Seifert, Jan Michálek,

Petr Kozel, Nela Nováková, and Tereza Holicová for collecting plant material and field assistance, and Salvatore Tomasello for help with data analysis. We thank John Bradican for his help with English editing of the manuscript. We thank four reviewers for their supportive comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.662715/full#supplementary-material

REFERENCES

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Argus, G. W. (2010). "Salix," in *Flora of North America*, Vol. 7 *Magnoliophyta: Salicaceae to Brassicaceae,* ed E. Flora of North America Editorial Committee (New York, NY: Oxford University Press), 23–51.

Azuma, T., Kajita, T., Yokoyama, J., and Ohashi, H. (2000). Phylogenetic relationships of Salix (Salicaceae) based on rbcL sequence data. *Am. J. Bot.* 87, 67–75. doi: 10.2307/2656686

Barcaccia, G., Meneghetti, S., Lucchin, M., and de Jong, H. (2014). Genetic segregation and genomic hybridization patterns support an allotetraploid structure and disomic inheritance for Salix species. *Diversity* 6, 633–651. doi: 10.3390/d6040633

Barrett, C. F., Specht, C. D., Leebens-Mack, J., Stevenson, D. W., Zomlefer, W. B., and Davis, J. I. (2014). Resolving ancient radiations: can complete plastid gene sets elucidate deep relationships among the tropical gingers (Zingiberales)? *Ann. Bot.* 113, 119–133. doi: 10.1093/aob/mct264

Besnard, G., Hernández, P., Khadari, B., Dorado, G., and Savolainen, V. (2011). Genomic profiling of plastid DNA variation in the Mediterranean olive tree. *BMC Plant Biol.* 11:80. doi: 10.1186/1471-2229-11-80

Bock, D. G., Andrew, R. L., and Rieseberg, L. H. (2014). On the adaptive value of cytoplasmic genomes in plants. *Mol. Ecol.* 23, 4899–4911. doi: 10.1111/mec.12920

Bock, R., and Timmis, J. N. (2008). Reconstructing evolution: gene transfer from plastids to the nucleus. *BioEssays* 30, 556–566. doi: 10.1002/bies.20761

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303. doi: 10.1093/genetics/134.4.1289

Chen, A. J., Sun, H., Wen, J., Yang, Y., Chen, J., Sun, H., et al. (2010). *Molecular Phylogeny of Salix L. (Salicaceae) Inferred From Three Chloroplast Datasets and Its Systematic Implications*. International Association for Plant Taxonomy (IAPT) Stable, 29–37. Available online at: http://www.jstor.org/stable/27757048

Chen, J. (2020). Characterization of the complete chloroplast genome of *Salix variegata* (Salicaceae). *Mitochondrial DNA Part B Resour.* 5, 196–197. doi: 10.1080/23802359.2019.1698989

Clement, M., Posada, D., and Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* 9, 1657–1659. doi: 10.1046/j.1365-294x.2000.01020.x

Dong, W., Liu, J., Yu, J., Wang, L., and Zhou, S. (2012). Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS ONE* 7:e35071. doi: 10.1371/journal.pone.0035071

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., et al. (2015). *ycf*1, the most promising plastid DNA barcode of land plants. *Sci. Rep.* 5:8348. doi: 10.1038/srep08348

Duvall, M. R., Burke, S. V., and Clark, D. C. (2020). Plastome phylogenomics of Poaceae: alternate topologies depend on alignment gaps. *Bot. J. Linn. Soc.* 192, 9–20. doi: 10.1093/botlinnean/boz060

Eaton, D. A. R., and Overcast, I. (2016). *iPYRAD: Interactive Assembly and Analysis of RADseq Data Sets*. Available online at: available: http://ipyrad.readthedocs.io/

Fang, C., Zhao, S., and Skvortsov, A. (1999). "Salicaceae," in *Flora of China, Vol. 4*, eds Z. Y. Wu and P. H. Raven (St. Louis, MI: Science Press, Beijing and Missouri Botanical Garden Press), 139–274.

Flot, J. F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E. G. J., et al. (2013). Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* 500, 453–457. doi: 10.1038/nature12326

Friedberg, E. C., and Meira, L. B. (2006). Database of mouse strains carrying targeted mutations in genes affecting biological responses to DNA damage Version 7. *DNA Repair* 5, 189–209. doi: 10.1016/j.dnarep.2005.09.009

Gaut, B., Yang, L., Takuno, S., and Eguiarte, L. E. (2011). The patterns and causes of variation in plant nucleotide substitution rates. *Annu. Rev. Ecol. Evol. Syst.* 42, 245–266. doi: 10.1146/annurev-ecolsys-102710-145119

Gitzendanner, M. A., Soltis, P. S., Wong, G. K. S., Ruhfel, B. R., and Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *Am. J. Bot.* 105, 291–301. doi: 10.1002/ajb2.1048

Givnish, T. J., Spalink, D., Ames, M., Lyon, S. P., Hunter, S. J., Zuluaga, A., et al. (2015). Orchid phylogenomics and multiple drivers of their extraordinary diversification. *Proc. R. Soc. B Biol. Sci.* 282, 171–180. doi: 10.1098/rspb.2015.1553

Gramlich, S., Wagner, N. D., and Hörandl, E. (2018). RAD-seq reveals genetic structure of the F2-generation of natural willow hybrids (*Salix* L.) and a great potential for interspecific introgression. *BMC Plant Biol.* 18:317. doi: 10.1186/s12870-018-1552-6

He, L., Jia, K. H., Zhang, R. G., Wang, Y., Shi, T., and Le, L.i, Z. C., et al. (2021a). Chromosome-scale assembly of the genome of *Salix dunnii* reveals a male-heterogametic sex determination system on chromosome 7. *Mol. Ecol. Resour.* 21, 1966–1982. doi: 10.1101/2020.10.09.333229

He, L., Wagner, N. D., and Hörandl, E. (2021b). Restriction-site associated DNA sequencing data reveal a radiation of willow species (*Salix* L., Salicaceae) in the Hengduan Mountains and adjacent areas. *J. Syst. Evol.* 59, 44–57. doi: 10.1111/jse.12593

Hörandl, E. (1992). Die Gattung *Salix* in Österreich (mit Berücksichtigung angrenzender Gebiete). *Abh. Zool.-Bot. Ges. Österreich* 27, 1–170.

Hörandl, E., Florineth, F., and Hadacek, F. (2012). *Weiden in Österreich und Angrenzenden Gebieten (Willows in Austria and Adjacent Regions)*, 2nd ed. Vienna: University of Agriculture.

Huang, D. I., Hefer, C. A., Kolosova, N., Douglas, C. J., and Cronk, Q. C. B. (2014). Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa (Salicaceae). New Phytol.* 204, 693–703. doi: 10.1111/nph.12956

Huang, Y., Wang, J., Yang, Y., Fan, C., and Chen, J. (2017). Phylogenomic analysis and dynamic evolution of chloroplast genomes in Salicaceae. *Front. Plant Sci.* 8, 1–13. doi: 10.3389/fpls.2017.01050

Jia, H., Wang, L., Li, J., Sun, P., Lu, M., and Hu, J. (2020). Physiological and metabolic responses of *Salix sinopurpurea* and *Salix suchowensis* to drought stress. *Trees Struct. Funct.* 34, 563–577. doi: 10.1007/s00468-019-01937-z

Karp, A., Hanley, S. J., Trybush, S. O., Macalpine, W., Pei, M., and Shield, I. (2011). Genetic improvement of willow for bioenergy and biofuels. *J. Integr. Plant Biol.* 53, 151–165. doi: 10.1111/j.1744-7909.2010.01015.x

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Khakhlova, O., and Bock, R. (2006). Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant J.* 46, 85–94. doi: 10.1111/j.1365-313X.2006.02673.x

Kikuchi, S., Bédard, J., Hirano, M., Hirabayashi, Y., Oishi, M., Imai, M., et al. (2013). Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* 339, 571–574. doi: 10.1126/science.1229262

Lauron-Moreau, A., Pitre, F. E., Argus, G. W., Labrecque, M., and Brouillet, L. (2015). Phylogenetic relationships of American willows (*Salix* L., Salicaceae). *PLoS ONE* 10:e0121965. doi: 10.1371/journal.pone.0121965

Leskinen, E., and Alström-Rapaport, C. (1999). Molecular phylogeny of Salicaceae and closely related Flacourtiaceae: evidence from 5.8 S, ITS 1 and ITS 2 of the rDNA. *Plant Syst. Evol.* 215, 209–227. doi: 10.1007/BF00984656

Li, M. M., Wang, D. Y., Zhang, L., Kang, M. H., Lu, Z. Q., Zhu, R., et al. (2019). Intergeneric relationships within the family Salicaceae s.l. based on plastid phylogenomics. *Int. J. Mol. Sci.* 20:3788. doi: 10.3390/ijms20153788

Lu, D., Hao, L., Huang, H., and Zhang, G. (2019). The complete chloroplast genome of *Salix psamaphila*, a desert shrub in northwest China. *Mitochondrial DNA Part B Resour.* 4, 3432–3433. doi: 10.1080/23802359.2019.1675485

McKain, M. R., and Wilson, M. (2017). Fast-Plast: Rapid *de novo* assembly and finishing for whole chloroplast genomes. Available online at: https://github.com/mrmckain/Fast-Plast

Mower, J. P., Stefanovi,ć, S., Hao, W., Gummow, J. S., Jain, K., Ahmed, D., et al. (2010). Horizontal acquisition of multiple mitochondrial genes from a parasitic plant followed by gene conversion with host mitochondrial genes. *BMC Biol.* 8:150. doi: 10.1186/1741-7007-8-150

Mu, X. Y., Tong, L., Sun, M., Zhu, Y. X., Wen, J., Lin, Q. W., et al. (2020). Phylogeny and divergence time estimation of the walnut family (Juglandaceae) based on nuclear RAD-Seq and chloroplast genome data. *Mol. Phylogenet. Evol.* 147:106802. doi: 10.1016/j.ympev.2020.106802

Narango, D. L., Tallamy, D. W., and Shropshire, K. J. (2020). Few keystone plant genera support the majority of *Lepidoptera* species. *Nat. Commun.* 11, 1–8. doi: 10.1038/s41467-020-19565-4

Nargar, K., Molina, S., Wagner, N., Nauheimer, L., Micheneau, C., and Clements, M. A. (2018). Australasian orchid diversification in time and space: molecular phylogenetic insights from the beard orchids (*Calochilus*, Diurideae). *Aust. Syst. Bot.* 31, 389–408. doi: 10.1071/SB18027

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, R., McGlinn, D., et al. (2019). *Vegan: Community Ecology Package*. R package version 2.5-6. Available online at: https://CRAN.Rproject.org/package=vegan

Palo, R. T. (1984). Distribution of birch (*Betula* SPP.), willow (*Salix* SPP.), and poplar (*Populus* SPP.) secondary metabolites and their potential role as chemical defense against herbivores. *J. Chem. Ecol.* 10, 499–520. doi: 10.1007/BF00988096

Paradis, E., and Schliep, K. (2019). Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Parks, M., Cronn, R., and Liston, A. (2012). Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). *BMC Evol. Biol.* 12:100. doi: 10.1186/1471-2148-12-100

Percy, D. M., Argus, G. W., Cronk, Q. C., Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., et al. (2014). Understanding the spectacular failure of DNA barcoding in willows (*Salix*): does this result from a trans-specific selective sweep? *Mol. Ecol.* 23, 4737–4756. doi: 10.1111/mec.12837

Pham, K. K., Hipp, A. L., Manos, P. S., and Cronn, R. C. (2017). A time and a place for everything: phylogenetic history and geography as joint predictors of oak plastome phylogeny. *Genome* 60, 720–732. doi: 10.1139/gen-2016-0191

Piatczak, E., Dybowska, M., Płuciennik, E., Kośla, K., Kolniak-ostek, J., and Kalinowska-lis, U. (2020). Identification and accumulation of phenolic compounds in the leaves and bark of *Salix alba* (L.) and their biological potential. *Biomolecules* 10, 1–17. doi: 10.3390/biom10101391

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Rambaut, A. (2014). *Figtree, A Graphical Viewer of Phylogenetic Trees*. Available online at: http://tree.bio.ed.ac.uk/software/figtree

Rieseberg, L. H., and Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic geneflow in plants. *Evol. Trends Plants* 5, 65–84.

Ripma, L. A., Simpson, M. G., and Hasenstab-Lehman, K. (2014). Geneious! Simplified genome skimming methods for phylogenetic systematic studies: a case study in *Oreocarya (Boraginaceae)*. *Appl. Plant Sci.* 2:1400062. doi: 10.3732/apps.1400062

Savage, J. A., and Cavender-Bares, J. (2012). Habitat specialization and the role of trait lability in structuring diverse willow (genus *Salix*) communities. *Ecology* 93, 138–150. doi: 10.1890/11-0406.1

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., et al. (2005). The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot.* 92, 142–166. doi: 10.3732/ajb.92.1.142

Shaw, J., Lickey, E. B., Schilling, E. E., and Small, R. L. (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the Tortoise and the hare III. *Am. J. Bot.* 94, 275–288. doi: 10.3732/ajb.94.3.275

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47, W65–W73. doi: 10.1093/nar/gkz345

Skvortsov, A. K. (1999). "Willows of Russia and adjacent countries," in *Taxonomical and Geographical Revision*, eds A. G. Zinovjev, G. W. Argus, J. Tahvanainen, and H. Roininen (Finland: Joensuu).

Smart, L. B., Volk, T. A., Lin, J., Kopp, R. F., Phillips, I. S., Cameron, K. D., et al. (2005). Genetic improvement of shrub willow (*Salix* spp.) crops for bioenergy and environmental applications in the United States. *Unasylva* 56, 51–55.

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

Straub, S. C. K., Moore, M. J., Soltis, P. S., Soltis, D. E., Liston, A., and Livshultz, T. (2014). Phylogenetic signal detection from an ancient rapid radiation: effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. *Mol. Phylogenet. Evol.* 80, 169–185. doi: 10.1016/j.ympev.2014.07.020

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., and Liston, A. (2012). Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *Am. J. Bot.* 99, 349–364. doi: 10.3732/ajb.1100335

Suda, Y., and Argus, G. W. (1968). Chromosome numbers of some North American *Salix*. *Brittonia* 20, 191–197. doi: 10.2307/2805440

Taberlet, P., Gielly, L., Pautou, G., and Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol. Biol.* 17, 1105–1109. doi: 10.1007/BF00037152

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. and Gray). *Science* 313, 1596–604. doi: 10.1126/science.1128691

Wagner, N. D., Gramlich, S., and Hörandl, E. (2018). RAD sequencing resolved phylogenetic relationships in European shrub willows (Salix L. subg. Chamaetia and subg. Vetrix) and revealed multiple evolution of dwarf shrubs. *Ecol. Evol.* 8, 8243–8255. doi: 10.1002/ece3.4360

Wagner, N. D., He, L., and Hörandl, E. (2020). Phylogenomic relationships and evolution of polyploid *Salix* species revealed by RAD sequencing data. *Front. Plant Sci.* 11, 1–38. doi: 10.3389/fpls.2020.01077

Wagner, N. D., He, L., and Hörandl, E. (2021). The evolutionary history, diversity, and ecology of willows (*Salix* l.) in the European alps. *Diversity* 13, 1–16. doi: 10.3390/d13040146

Wicke, S., and Schneeweiss, G. M. (2015). "Next-generation organellar genomics: potentials and pitfalls of highthroughput technologies for molecular evolutionary studies and plant systematics," in *Next Generation Sequencing in Plant Systematics. Regnum; Vegetabile Book Series of the IAPT (International Association of Plant Taxonomy),* eds E. Hörandl and M. Appelhans (Königstein: Koeltz Scientific Books). doi: 10.14630/000002

Wolfe, A. D., and Randle, C. P. (2004). Recombination, heteroplasmy, haplotype polymorphism, and paralogy in plastid genes: implications for plant molecular systematics. *Syst. Bot.* 29, 1011–1020. doi: 10.1600/0363644042451008

Wolfe, K. H., Li, W. H., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U.S.A.* 84, 9054–9058. doi: 10.1073/pnas.84.24.9054

Wu, D., Wang, Y., and Zhang, L. (2019). The complete chloroplast genome sequence of an economic plant *Salix wilsonii*. *Mitochondrial DNA Part B* 4, 3560–3562. doi: 10.1080/23802359.2019.1668311

Wu, J., Nyman, T., Wang, D.-C., Argus, G. W., Yang, Y.-P., and Chen, J.-H. (2015). Phylogeny of Salix subgenus *Salix* s.l. (Salicaceae): delimitation, biogeography, and reticulate evolution. *BMC Evol. Biol.* 15:31. doi: 10.1186/s12862-015-0311-7

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Zhang, L., Xi, Z., Wang, M., Guo, X., and Ma, T. (2018). Plastome phylogeny and lineage diversification of Salicaceae with focus on poplars and willows. *Int. J. Bus. Innov. Res.* 17, 7817–7823. doi: 10.1002/ece3.4261

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer RS declared a shared affiliation, with no collaboration, with one of the authors, MV, to the handling editor at the time of the review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 15 October 2021
doi: 10.3389/fpls.2021.725900

Phylogenomics of *Salvia* L. subgenus *Calosphace* (Lamiaceae)

Sabina Irene Lara-Cabrera[1†], Maria de la Luz Perez-Garcia[2†], Carlos Alonso Maya-Lastra[3†], Juan Carlos Montero-Castro[1†], Grant T. Godden[4], Angelica Cibrian-Jaramillo[5†], Amanda E. Fisher[6†] and J. Mark Porter[7]*

[1] *Laboratorio de Sistemática Molecular de Plantas, Facultad de Biología, Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Mexico,* [2] *Departamento de Botánica y Zoología, Centro Universitario de Ciencias Biológicas y Agropecuarias, Universidad de Guadalajara, Guadalajara, Mexico,* [3] *Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY, United States,* [4] *Florida Museum of Natural History, University of Florida, Gainesville, FL, United States,* [5] *Laboratorio Nacional de Genómica para la Biodiversidad, Unidad de Genómica Avanzada del Centro de Investigación y de Estudios Avanzados del instituto Politécnico Nacional, Irapuato, Mexico,* [6] *Department of Biological Sciences, California State University, Long Beach, CA, United States,* [7] *California Botanic Garden, Claremont, CA, United States*

OPEN ACCESS

Edited by:
Stefan Wanke, Technische Universität Dresden, Germany

Reviewed by:
Aaron Liston, Oregon State University, United States
Roswitha Schmickl, Academy of Sciences of the Czech Republic (ASCR), Czechia

***Correspondence:**
Sabina Irene Lara-Cabrera
sabina.lara@umich.mx; slaracabrera@gmail.com

†ORCID:
Sabina Irene Lara-Cabrera
orcid.org/0000-0001-8551-9829
Maria de la Luz Perez-Garcia
orcid.org/0000-0001-5272-5052
Carlos Alonso Maya-Lastra
orcid.org/0000-0002-0550-3331
Juan Carlos Montero Castro
orcid.org/0000-0002-3098-14150
Angélica Cibrián-Jaramillo
orcid.org/000-002-7974-455X
Amanda E. Fisher
orcid.org/0000-0002-9928-9558

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *16 June 2021*
Accepted: *07 September 2021*
Published: *15 October 2021*

Citation:
Lara-Cabrera SI, Perez-Garcia MdlL, Maya-Lastra CA, Montero-Castro JC, Godden GT, Cibrian-Jaramillo A, Fisher AE and Porter JM (2021) Phylogenomics of Salvia L. subgenus Calosphace (Lamiaceae). Front. Plant Sci. 12:725900. doi: 10.3389/fpls.2021.725900

The evolutionary relationships of *Salvia* have been difficult to estimate. In this study, we used the Next Generation Sequencing method Hyb-Seq to evaluate relationships among 90 Lamiaceae samples, including representatives of *Mentheae*, *Ocimeae*, *Salvia* subgenera *Audibertia*, *Leonia*, *Salvia*, and 69 species of subgenus *Calosphace*, representing 32 of Epling's sections. A bait set was designed in MarkerMiner using available transcriptome data to enrich 119 variable nuclear loci. Nuclear and chloroplast loci were assembled with *hybphylomaker* (HPM), followed by coalescent approach analyses for nuclear data (ASTRAL, BEAST) and a concatenated Maximum Likelihood analysis of chloroplast loci. The HPM assembly had an average of 1,314,368 mapped reads for the sample and 527 putative exons. Phylogenetic inferences resolved strongly supported relationships for the deep-level nodes, agreeing with previous hypotheses which assumed that subgenus *Audibertia* is sister to subgenus *Calosphace*. Within subgenus *Calosphace,* we recovered eight monophyletic sections *sensu* Epling, *Cardinalis*, *Hastatae*, *Incarnatae,* and *Uricae* in all the analyses (nDNA and cpDNA), *Biflorae*, *Lavanduloideae,* and *Sigmoideae* in nuclear analyses (ASTRAL, BEAST) and *Curtiflorae* in ASTRAL trees. Network analysis supports deep node relationships, some of the main clades, and recovers reticulation within the core *Calosphace*. The chloroplast phylogeny resolved deep nodes and four monophyletic *Calosphace* sections. Placement of *S. axillaris* is distinct in nuclear evidence and chloroplast, as sister to the rest of the *S*. subg. *Calosphace* in chloroplast and a clade with "*Hastatae* clade" sister to the rest of the subgenus in nuclear evidence. We also tested the monophyly of *S. hispanica*, *S. polystachia*, *S. purpurea*, and *S. tiliifolia,* including two samples of each, and found that *S. hispanica* and *S. purpurea* are monophyletic. Our baits can be used in future studies of Lamiaceae phylogeny to estimate relationships between genera and among species. In this study, we presented a Hyb-Seq phylogeny for complex, recently diverged *Salvia*, which could be implemented in other Lamiaceae.

Keywords: Hyb-Seq, chloroplast, section, nuclear, monophyly

INTRODUCTION

Phylogenetic relationships for many plant groups have been studied through the last 30–40 years at deep (APG, 1998; Zeng et al., 2017; Breinholt et al., 2021) and shallow phylogenetic levels (Wells et al., 2020), mostly through Sanger sequencing (Sanger et al., 1977) and recently through Next Generation Sequencing (Wanke et al., 2017; Carlsen et al., 2018; Herrando-Moraira and The Cardueae Radiations Group, 2018; Villaverde et al., 2018; Carter et al., 2019; Johnson et al., 2019). However, in groups with recent radiation events (Larridon et al., 2020) such as *Salvia* L. (Walker and Sytsma, 2007; Jenks et al., 2013; Fragoso-Martínez et al., 2018; González-Gallegos et al., in press), many questions remain at the shallow-phylogenetic scale, such as relationships among sections, among species, and species monophyly.

The sages (*Salvia*) with *ca*. 1,000 species (Harley et al., 2004; Drew et al., 2017), are among the largest angiosperm genera (Frodin, 2004). They are widely distributed with many economically important species (Wu et al., 2012; Lopresti, 2017). *Salvia* flowers are bilabiate and have evolved a wide variety of showy colors and shapes (Lara-Cabrera et al., in press), as well as staminal levers and other morphological adaptations to pollinators (Claßen-Bockhoff et al., 2004; Wester and Claßen-Bockhoff, 2011; Benítez-Vieyra et al., 2014; Kriebel et al., 2019, 2020; Celep et al., 2020). Previous *Salvia* phylogenies that employed few, e.g., <5–10, chloroplast or nuclear coding and non-coding loci were successful in reconstructing relationships at many deep-level nodes. These studies showed that *Salvia* is polyphyletic with five embedded genera, namely, *Dorystaechas* Boiss. and Heldr. ex Benth., *Meriandra* Benth., *Perovskia* Kar., *Rosmarinus* L., and *Zhumeria* Rech. f. and Wendelbo (Walker et al., 2004; Walker, 2006; Walker and Sytsma, 2007). *Salvia* species are classified into five subgenera, namely, *Salvia*, *Audibertia* J. B. Walker, B. T. Drew and K. J. Sytsma, *Calosphace* (Benth.) Epling, *Leonia* Cerv., and *Sclarea* Mill. A proposal to "lump" these genera into *Salvia* would add five more subgenera to *Salvia* (Drew et al., 2017), which are *Dorystaechas* (Boiss. and Heldr. ex Benth.) J. B.Walker, B. T. Drew, and J. G. González, *Meriandra* (Benth.); J. B. Walker, B. T. Drew, and J. G. González, *Perovskia* (Kar.); J. B. Walker, B. T. Drew, and J. G. González, *Rosmarinus* (L.); J. B. Walker, B. T. Drew, and J. G. González, and *Zhumeria* (Rech.f. and Wendelbo); J. B. Walker, B. T. Drew, and J. G. González. Among these, we focused in this study mainly on the American subgenus *Calosphace* and some representatives in subgenera *Audibertia, Leonia*, and *Salvia s.s.*

Salvia subg. *Calosphace* is distributed from southern USA to Argentina (Ramamoorthy and Elliott, 1998; Walker et al., 2004), with *ca*. 580 (González-Gallegos et al., in press) to 600 species (Martínez-Gordillo et al., 2017). It is most diverse in Mexico and Central America (275 species), the Andes (155 species), Eastern South America (60 species), and the Antilles (45 species; Jenks et al., 2013). Given *S*. subg. *Calosphace* species diversity and morphological complexities, it has been classified into 102 sections (Epling, 1939, 1940, 1941, 1944, 1947, 1951; Epling and Mathias, 1957; Epling and Jativa, 1963). However, the sectional classification has been criticized (Standley and Williams, 1973; Torke, 2000; Walker, 2006; Wood, 2007), given the few characters employed to define sections, and disjunct distribution of some species. Regardless, Epling's classification is recognized as a necessary starting point to further the study on *Salvia* until a new monograph is compiled (Ramamoorthy, 1984; Wood, 2007; Klitgaard, 2012).

Previous phylogenetic studies of *Calosphace* resolved *S. axillaris* Moc. and Sessé sister to the rest of the subgenus (Walker et al., 2004; Walker and Sytsma, 2007; Jenks et al., 2013; Drew et al., 2017; Fragoso-Martínez et al., 2018; Kriebel et al., 2019), followed by the *Hastatae* clade (*Salvia patens* Ort. + *Salvia vitifolia* Benth.); members of the *S*. sects. *Tomentellae*, *Dusenostachys*, *Uliginosae*, *Erytrostachys*, *Micranthae*, *Fulgentes,* and *Membranaceae* (Fragoso-Martínez et al., 2018) or *Fulgentes* was paraphyletic to members of sects. *Cardinalis* and *Flocculosae* (Jenks et al., 2013). The "Core *Calosphace*" contains the most species and relationships within this clade that have been difficult to resolve or have had low branch support. The "Core *Calosphace*" clade was initially described by Walker (2006) and refers to a clade of "core radiation" that is "difficult to characterize morphologically but is well-supported in the molecular analyses...". It has been hypothesized that recent divergence events are clouding the phylogenetic signal, which could be further tested with expanded taxon sampling and additional phylogenetically informative sequence data (Olvera-Mendoza et al., 2020; Villaverde et al., 2020). This was attempted by Fragoso-Martínez et al. (2017) and Kriebel et al. (2019) using hybrid enrichment protocols across *Salvia* and to test sectional monophyly of the *Calosphace*. The Anchored Hybrid Enrichment (AHE) Angiosperm kit v. 1 (Buddenhagen et al., 2016) was tested on 12 *Salvia* species and captured 399 nuclear loci (Fragoso-Martínez et al., 2017) and later the protocol was used for 35 *Salvia* (13 *Calosphace* and 2 *Audibertia*) species capturing 316 nuclear genes (Kriebel et al., 2019). Both phylogenies improved clade resolution as compared to previous sequencing studies (Walker et al., 2004; Walker and Sytsma, 2007; Jenks et al., 2013; Will and Claßen-Bockhoff, 2017; Fragoso-Martínez et al., 2018; Hu et al., 2018).

In this study, we used the Hyb-Seq protocol (Weitemier et al., 2014) for target enrichment of low copy nuclear exons and flanking regions and genome skimming of organellar genomes. Hyb-Seq has been successfully used to solve shallow-level phylogenetic relationships in *Asclepias* L. (Straub et al., 2011, 2012), Annonaceae (Couvreur et al., 2019), *Asteraceae* (Mandel et al., 2017; Herrando-Moraira and The Cardueae Radiations Group, 2019; Johnson et al., 2019; Jones et al., 2019), Poaceae (Fisher et al., 2016), and *Rubus* (Carter et al., 2019), among others. We used MarkerMiner (Chamala et al., 2015) to identify low copy nuclear loci in 22 Lamiaceae transcriptomes (including *Salvia officinalis* L. and *S. splendens* Sellow ex Schult.) and design both general and specific purpose bait sets. We sampled a total of 90 Lamiaceae from tribes Mentheae and Ocimeae, 75 samples represent 32 of Epling's *S*. subg. *Calosphace* sections. Our goals were to test classification of Epling and relationships found in previous studies of subg. *Calosphace*; test species monophyly for four important and morphologically

complex species. Furthermore, we aimed to identify sufficiently polymorphic loci for future studies in *Salvia*.

METHODS

Taxonomic Sampling

The study materials consisted of 90 *Lamiaceae* from nine genera which were sampled (**Supplementary Table 1**). Exactly 10 species were sampled from tribe *Mentheae* [*Agastache pallidiflora* subsp. *neomexicana* (Briq.) Lint and Epling, *Dracocephalum parviflorum* Nutt., *Hedeoma drummondii* Benth., *Lepechinia hastata* (A. Gray) Epling, *Lepechinia* sp., *Lycopus americanus* Muhl., *Melissa officinalis* L., *Poliomintha incana* (Torr.) A. Gray, and *Prunella vulgaris* L.] and one species was sampled from the tribe *Ocimeae* [*Cantinoa mutabilis* (Rich.) Harley and J. F. B. Pastore] to root the trees (Li et al., 2016).

Multiple subgenera of *Salvia* were represented in our sampling, two each from the *S.* subg. *Audibertia* sect. *Audibertia* (*S. brandegeei* Munz and *S. sonomensis* Greene) and *S.* subg. *Salvia* sect. *Salvia* (the Mediterranean *S. officinalis* and the Malagasy *S. sessilifolia* A. Gray ex S. Watson), and one from the *S.* subg. *Leonia* sect. *Salviastrum* [S. *texana* (Scheele) Torr.]. From the *S.* subgenus *Calosphace,* we sampled 72 species (**Supplementary Table 1**) in all, representing 32 of the 102 sections *sensu* Epling. Our sampling represents the geographic range of the taxon in Mexico (67 species; **Supplementary Table 1**) and includes five additional species from Central and South America (*S. pauciserrata* Benth., *S. scutellarioides* Kunth, *S. splendens*, *S. squalens* Kunth, and *S. tubiflora* Sm.). Seven species were sampled for molecular study for the first time (*S. brachyodonta* Briq., *S. decora* Epling, *S. dichlamys* Epling, *S. perblanda* Epling, *S. puberula* Fernald, *S. purepecha* Bedolla, S. Lara Cabrera and Zamudio, and *S. roscida* Fernald). Additionally, we included two samples from distinct provenances for *Salvia hispanica* L., *Salvia polystachia* Cav., *Salvia purpurea* Cav., and *Salvia tiliifolia* Vahl., to assess their monophyly, which further tested the resolving power of this protocol.

Phylogenetic Marker Selection, Bait Design, and DNA Sequencing

Genomic DNA was isolated from 10 mg of silica-dried leaf material using a modified 2X CTAB protocol (Doyle and Doyle, 1987). DNAs were quantified using a Qubit 2.0 fluorometer (Thermo Fisher Scientific Inc., Waltham, MA, USA) and diluted to a concentration of 20 ng/μl. Afterward, 60 μl of DNA solution were transferred to a 96-well plate and shipped to Rapid Genomics (Gainesville, FL, USA) for library preparation, hybrid enrichment of nuclear loci, and paired-end (2 × 150 bp) sequencing on the Illumina HiSeq 2500 instrument (Thermo Fisher Scientific Inc.).

A multipurpose bait set was designed for use across independent research projects with *Salvia*, Acanthaceae, Clusiaceae, Lamiales, and Polemoniaceae. To select loci and provide sequence data for bait design for the *Salvia* and Lamiales studies, we analyzed a set of 77 transcriptomes from the One Thousand Plant Transcriptomes Initiative (OneKp), including 68 from Lamiales and 9 from outgroup taxa representing Boraginales, Gentianales, and Solanales (One Thousand Plant Transcriptomes Intitiative, 2019), and an additional transcriptome for *S. splendens* Sellow ex Wied-Neuw. in Genbank [Ge et al., 2014; https://www.ncbi.nlm.nih.gov/bioproject/422035 (Taxonomy ID: 180675)]. We used the MarkerMiner 1.0 (Chamala et al., 2015) pipeline with its default settings to assess putative orthology among transcripts in our data set with a set of *Arabidopsis thaliana* (L.) Heynh. transcripts from genes that were identified as single- (or low-)copy across angiosperms by an orthology analysis of 20 genomes (De Smet et al., 2013), mapping to chromosomes 1, 2, 3, 4, and 5 in the *A. thaliana* genome (**Table 1**); at the time we had no fully annotated Lamiaceae genome. Gene clusters identified by MarkerMiner were aligned with MAFFT (Katoh et al., 2002) and individually reviewed for marker selection.

The final selection of loci for bait design was based on the following criteria: sequence variability, align-ability, demonstrated phylogenetic utility within Lamiales and Lamiaceae (Godden, unpublished data), and economic considerations. The latter criterion dictated the numbers of loci and baits per project that could be accommodated in the final multipurpose bait set. Overall, baits in the multipurpose set relevant to this project included the following: 883 Lamiales general-purpose baits (76,272 bp) and 1,207 *Salvia*-specific baits (131,394 bp), based on the Lamiales transcriptomes and *S. officinalis* and *S. splendens* alignments for the latter (**Supplementary Table 2**). Paired baits were manufactured with TruSeq technology by myBaits (Daicel Arbor Biosciences, Ann Arbor, MI, USA). Samples were sequenced on an Illumina® HiSeq 2,500 as 150 bp PE reads. Raw read quality was assessed with Fastqc v.0.11.2 (Andrews, 2010; Babraham Bioinformatics, Cambridge, England). Adapter sequences and low-quality bases were trimmed using Cutadapt v. 1.8.1 (Martin, 2011).

Assembly

Raw reads were processed in HybPhyloMaker (HPM) v.1.6.4 (Fér and Schmickl, 2018), this pipeline contains multiple steps or scripts that allow assembly and further analyses (from here on throughout the text, these are quoted per acronym and numbered as specified in the script name from the HPM reference manual). Using the script HPM_0b in the pipeline, individual reads were mapped to two pseudo reference sequences. The first nuclear pseudo reference was the alignment of the probe set containing 527 putative exons (these were previously used as probes to target the specified genes) and the second pseudo reference was 114 chloroplast loci from *Salvia miltiorrhiza* Bunge complete plastome JX312195 (Qian et al., 2013), separated by 400 Ns to capture any chloroplast sequences.

In order to summarize the effectiveness of capture based on our nuclear pseudo reference, we used all sequences for each exon produced by HPM_3 and calculated the missing data for each of them compared with the original probes in a heat map (**Figure 1**).

The reads were trimmed, filtered, and mapped to create the alignments for reconstructing gene and species trees, using the following steps: script HPM_1 was used to remove sequencing adapters and trim reads based on their quality using

TABLE 1 | HPM assembly characteristics per sample for 90 samples targeting 119 nuclear genes, 26 genes were later filtered through the next steps in HPM.

Species	Total nr. reads	Nr. paired reads	Nr. forward unpaired reads	Nr. reverse unpaired reads	Nr. mapped reads	% Mapped reads
Agastache sp.	4,795,650	2,341,728	68,327	41,718	1,328,625	27.70
Dracocephalum parviflorum	3,520,197	1,720,163	48,401	30,827	1,314,368	37.34
Hedeoma drummondii	3,279,317	1,590,347	63,879	34,020	1,099,519	33.53
Cantinoa mutabilis	2,132,204	1,028,383	50,662	24,360	606,056	28.42
Lepechinia hastata	1,769,299	859,751	31,856	17,489	343,698	19.43
Lepechinia sp.	2,195,750	1,071,781	32,194	19,383	589,054	26.83
Lycopus americanus	4,277,985	2,070,253	92,390	42,301	1,518,980	35.51
Melissa officinalis	3,152,455	1,540,899	44,739	25,608	898,696	28.51
Poliomintha incana	3,122,237	1,521,007	47,608	32,011	646,254	20.70
Prunella vulgaris	1,390,762	676,453	24,947	12,281	346,864	24.94
Salvia aequidistans	3,916,050	1,893,614	89,280	38,855	1,372,874	35.06
Salvia amarissima	2,304,666	1,110,067	51,892	32,269	745,462	32.35
Salvia areolata	4,393,342	2,139,596	66,078	46,336	1,772,640	40.35
Salvia axillaris	2,181,941	1,049,876	45,833	35,955	636,302	29.16
Salvia azurea	4,533,709	2,206,512	70,640	48,461	1,751,642	38.64
Salvia blepharophylla	4,815,199	2,345,302	74,133	47,050	2,001,549	41.57
Salvia brachyodonta	4,295,551	2,082,412	83,182	46,634	1,559,488	36.30
Salvia brandegeei	2,312,395	1,125,113	39,632	21,346	921,779	39.86
Salvia breviflora	853,293	409,465	19,377	14,703	**255,228**	29.91
Salvia cacaliifolia	3,228,987	1,571,229	55,214	30,241	1,027,027	31.81
Salvia chamaedryoides	4,695,712	2,269,098	127,848	28,614	1,911,882	40.72
Salvia chiapensis	2,079,209	1,002,758	46,707	26,286	648,845	31.21
Salvia cinnabarina	3,864,253	1,800,499	240,431	20,724	1,098,123	28.42
Salvia clinopodioides	2,663,514	1,230,419	186,821	13,104	640,127	24.03
Salvia coahuilensis	6,760,892	3,139,087	448,389	30,673	2,648,399	39.17
Salvia connivens	3,955,980	1,837,482	257,520	21,352	1,614,082	40.80
Salvia curtiflora	4,897,789	2,272,279	326,202	23,555	2,001,741	40.87
Salvia curviflora	1,480,065	654,763	158,472	11,637	414,403	28.00
Salvia decora	2,188,709	941,220	291,136	14,110	723,442	33.05
Salvia dichlamys	4,118,055	1,911,522	270,852	21,762	1,667,989	40.50
Salvia disjuncta	3,907,516	1,790,553	304,076	18,348	1,614,927	41.33
Salvia divinorum	3,160,115	1,443,708	257,368	14,228	853,477	27.01
Salvia dugesii	1,475,191	641,769	182,077	8,698	402,528	27.29
Salvia elegans	5,762,916	2,668,946	393,790	26,297	1,966,249	34.12
Salvia farinácea	4,399,124	1,997,356	384,195	18,243	1,523,587	34.63
Salvia filipes	5,580,520	2,608,588	334,714	26,475	2,122,980	38.04
Salvia fulgens	3,937,194	1,830,795	255,099	18,437	1,456,024	36.98
Salvia gesneriiflora	3,283,816	1,539,337	183,426	19,454	1,304,375	39.72
Salvia greggii	4,790,649	2,224,569	316,459	22,633	1,754,437	36.62
Salvia helianthemifolia	5,422,154	2,532,139	327,927	26,099	2,300,789	42.43
Salvia hispanica [10,685]	3,630,464	1,757,205	76,245	38,871	1,210,285	33.34
Salvia hispanica [16]	866,851	373,743	111,831	7,281	258,323	29.80
Salvia inconspicua	4,180,128	1,934,091	277,441	33,209	1,796,346	42.97
Salvia involucrata	3,456,571	1,590,456	257,524	16,111	1,201,438	34.76
Salvia iodantha	4,074,414	1,892,701	266,532	20,305	1,743,805	42.80
Salvia karwinskii	1,894,444	840,996	199,183	12,635	495,679	26.16
Salvia keerlii	4,646,169	2,111,991	396,028	23,818	1,656,324	35.65
Salvia lasiantha	3,372,169	1,561,766	215,150	31,579	1,350,572	40.05
Salvia lavanduloides	3,198,116	1,458,535	262,745	16,701	978,455	30.59
Salvia leucantha	6,810,955	3,144,365	489,051	30,542	**2,868,385**	42.11

(Continued)

TABLE 1 | Continued

Species	Total nr. reads	Nr. paired reads	Nr. forward unpaired reads	Nr. reverse unpaired reads	Nr. mapped reads	% Mapped reads
Salvia longispicata	2,133,006	960,688	195,062	15,867	550,342	25.80
Salvia longistyla	5,501,252	2,548,398	372,042	27,520	2,141,218	38.92
Salvia macrophylla	4,280,359	1,972,302	312,911	17,941	1,766,499	41.27
Salvia madrensis	3,028,391	1,353,831	306,003	13,126	954,658	31.52
Salvia melissodora	4,220,056	1,950,703	282,399	33,997	1,636,602	38.78
Salvia mexicana	4,828,203	2,230,598	341,298	23,015	1,985,539	41.12
Salvia microphylla	5,107,818	2,353,701	372,947	24,491	1,929,890	37.78
Salvia nepetoides	3,227,624	1,474,283	242,157	35,970	983,117	30.46
Salvia nervata	3,113,755	1,413,887	264,897	18,405	1,087,402	34.92
Salvia occidua	1,722,967	773,637	155,732	19,713	453,608	26.33
Salvia officinalis	3,898,073	1,787,315	293,750	19,578	1,582,911	40.61
Salvia patens	2,250,418	1,040,462	155,325	11,942	627,819	27.90
Salvia pauciserrata	4,359,118	2,003,658	327,793	20,915	1,831,230	42.01
Salvia perblanda	2,689,851	1,190,487	293,490	13,787	1,032,314	38.38
Salvia plurispicata	4,256,275	1,967,417	297,047	22,526	1,789,752	42.05
Salvia polystachia [163]	3,825,101	1,773,511	254,027	22,340	1,495,285	39.09
Salvia polystachia [065]	6,430,975	2,967,587	461,330	31,384	2,757,881	42.88
Salvia puberula	2,241,725	979,953	267,594	13,313	562,436	25.09
Salvia purépecha	3,088,154	1,389,198	283,125	25,687	898,212	29.09
Salvia purpurea [103]	4,165,410	1,903,329	334,825	21,969	1,547,632	37.15
Salvia purpurea [156]	4,619,949	2,074,882	309,978	158,190	1,919,193	41.54
Salvia ramosa	6,793,863	3,096,212	545,286	49,591	2,624,801	38.63
Salvia regla	2,864,486	1,298,276	254,203	11,585	904,405	31.57
Salvia rhyacophila	1,073,159	512,149	30,083	18,428	282,069	26.28
Salvia roscida	1,818,209	771,331	261,587	13,090	405,068	22.28
Salvia scutellarioides	4,329,068	1,999,664	304,019	21,116	1,583,847	36.59
Salvia semiatrata	4,747,390	2,182,186	342,749	38,185	1,681,244	35.41
Salvia sessilifolia	5,376,946	2,465,483	408,983	26,462	2,137,458	39.75
Salvia sonomensis	3,102,641	1,433,203	218,542	14,254	1,224,534	39.47
Salvia splendens	6,315,864	2 945 936	386,705	31,866	2,095,385	33.18
Salvia squalens	6,709,903	3,086,040	499,774	28,806	2,673,573	39.85
Salvia texana	2,373,414	1,095,103	166,950	14,986	355,513	14.98
Salvia tiliifolia [5]	2,533,486	1,226,296	44,353	35,889	879,475	34.71
Salvia tiliifolia [15]	4,806,266	2,226,329	321,326	30,078	1,673,090	34.81
Salvia tonaticensis	5,311,730	2,473,768	333,255	28,388	2,018,740	38.01
Salvia tubiflora	6,348,744	2,916,775	480,536	28,358	2,464,460	38.82
Salvia univerticillata	3,756,401	1,717,023	300,936	19,717	1,033,316	27.51
Salvia urica	5,125,397	2,354,572	371,226	43,179	1,878,912	36.66
Salvia vitifolia	3,131,504	1,452,729	205,439	18,261	903,035	28.84
Salvia wagneriana	5,357,533	2,609,669	91,101	45,676	1,805,424	33.70
Total	337,889,127	157,329,258	20,636,288	2,393,220	1,328,625	3,085
Average	**3,754,324**	1,748,103	229,292	26,591	**1,314,368**	34.28

Bold font indicate highest and lowest Nr. mapped reads.

Trimmomatic v.0.33 (Bolger et al., 2014). All reads <Q20 were discarded, and the remaining reads were trimmed if the average quality in a 5 bp window was <Q20. Reads shorter than 36 bp were removed. In addition, HPM uses FastUniq v.1.1 (Xu et al., 2012) to remove duplicate reads. The script HPM_2 was used to map the quality filtered and trimmed reads to the baits pseudo reference using BWA v.0.7.16a (Li and Durbin, 2009). Mapped reads for each taxon were summarized with a consensus sequence using Kindel v.0.1.4 (Constantinides and Robertson, 2017) included in the HP pipeline. This used a 51% majority consensus rule to call bases and convert any base with low coverage (2x) to an uninformative base (N). This was

repeated to consecutively map the filtered reads to the chloroplast pseudo reference.

Consensus sequences were matched to sequences of target exons using BLAT v.35 (Kent, 2002) (https://www.ncbi.nlm.nih.gov/pubmed/11932250), with 90% similarity for all samples to produce PSLX files using the script HPM_3. The script "assembled_exons_to_fastas.py" (Weitemier et al., 2014) is used in the script HPM_4a to construct matrices for multiple alignments and add Ns for taxa that lack a particular exon. Also, with the script HPM_4a, sequences were aligned in MAFFT v. 7.305 (Katoh and Standley, 2013) and nuclear exons belonging to the same gene were concatenated using AMAS (Borowiec, 2016). Finally using the script HPM_5 taxa, we took a conservative approach and removed exons from the alignment if more than 70% of the sequence missing and if exons were recovered in fewer than 75% of the taxa. We also tested the effect of this approach by varying our criterion to 30, 50, and 75% missing data for loci shared by all species in the HPM_5 matrix.

The two resulting data sets comprised 119 targeted nuclear genes and 114 loci for the chloroplast. Both data sets were independently filtered as described above to remove genes from the alignment with excessive missing data. After filtering, the alignments included 96 nuclear genes and 114 chloroplast loci.

Phylogenetic Analyses

Bayesian and Maximum Likelihood (ML) multispecies coalescent-based approaches were used to reconstruct species trees for the nuclear data. For Bayesian inference, we used BEAST v. 2.5.2 (Bouckaert et al., 2019) for the genes obtained from the HPM pipeline. First, the best fitting molecular evolution model was obtained for each independent gene using jModelTest v. 2.1.10 (Darriba et al., 2012). Four models were selected as best fitting (GTR + G, HK + G, K80 + G, and SYM + G). We ran BEAUTI v. 2.5.2 (Bouckaert et al., 2014) using the template for StarBEAST to prepare the BEAST analysis input file. In the analysis, trees were unlinked and the strict clock model was used for all of them. Genes with the same molecular evolution model had linked parameters. Finally, a coalescent constant population model was used as a *prior* on the species tree. We ran BEAST for 1.6 B states, sampling every 5,000 states. Tracer v. 1.6 (Drummond and Rambaut, 2007) was used to check ESS values. To construct a maximum clade credibility tree, we used TreeAnnotator v. 2.5.6 (Bouckaert et al., 2019) setting a burn-in of 25% of the states and "Mean Height" for node heights.

For ML inference, we used the scripts HPM_6b and HPM_7, that execute FastTree 2.1.10 SSE3 (Price et al., 2010) using default parameters, to generate trees for every gene in our dataset and root them using the external group (*Cantinoa*). Next, the species tree was inferred using the coalescent-based approach implemented in ASTRAL-III v. 5.6.1 (Zhang et al., 2018) running the script HPM_8a with default parameters. To reconstruct the phylogenetic network, we used the 96 gene trees produced by HPM_7 as input to NANUQ (Allman et al., 2019) incorporated in the MSCquartets package (Rhodes et al., 2021) for R (R Core Team, 2017, Vienna, Austria). We set an alpha of 1e-5 and a beta of 0.95 with the goal of testing for a signal of network cycles in the quartets. Later, we used SplitsTree (Huson and Bryant, 2006) to plot the network using default parameters.

To test the robustness of the phylogenetic inferences obtained for both nDNA and cpDNA matrices, we compared trees with different percentages of missing data (30, 50, and 75% missing), and a tree that maintains loci for all the samples (as opposed to removing loci present in fewer than 75% of taxa). For each dataset with different missing data, we re-ran the nuclear ASTRAL reconstruction and the chloroplast FastTree analysis with the parameters described earlier.

RESULTS

Bait Success and Assembly

After removing low-quality sequences and loci with many missing taxa in HPM, 96 of 119 genes targeted by our respective bait sets were retained for analysis. Samples had an average of 1,314,368 mapped reads (**Table 1**), with the fewest in *S. breviflora* Moc. and Sesse ex Benth. (255,228 reads) and the most in *S. leucantha* Cav. (2,868,385 reads). The length of nuclear gene alignments ranged from 154 bp (AT1G05350) to 3,336 bp (AT4G19490). In total, 527 putative exons were recovered. However, about a third of the targeted exons were retained for further analysis (**Supplementary Tables 3**, **4**). The filtering step in HPM removed some of the 527 putative exons, given that exon capture was not homogeneous across all samples nor loci. Fewer base pairs were recovered for the outgroup than the in-group and the highest recovery was in *S. officinalis*, one of the transcriptomes used to design the *Salvia* baits.

The HPM chloroplast assembly for all 90 samples, using the *S. miltiorrhiza* genome (Qian et al., 2013) as a pseudoreference, recovered 75 CDS (59 in the LSC, 5 IR-B, 10 SSC, 1 IR-A), 29 tRNA (20 LSC, 7 IR-B, 1 SSC), 5 genes with introns (3 LSC, 1 IR-B, 1 SSC), 4 rRNA in the IR-B and 1 IGS in the LSC region (**Supplementary Table 5**); ranging in length from 36 bp (*rps*19) to 6,870 bp (*ycf* 2).

Phylogenetic Inferences

All nuclear phylogenetic inferences, with both coalescent analyses HPM [BEAST (**Figure 2**) and ASTRAL (**Supplementary Figure 1**)] recovered similar tree topologies, with some differences in shallow-level relationships. A network of the nuclear alignment (**Figure 3**) revealed the same groupings in the outgroup and some reticulation within the core *Calosphace* as we recovered in our phylogenetic analyses. A quartet hypothesis test showed that a majority of quartets had a tree-like signal, with only a few quartets better represented as four-cycle networks (**Figure 3**). We also tested if varying the missing data to 30 (**Supplementary Figure 2A**), 50 (**Supplementary Figure 2B**), or 70% (**Supplementary Figure 2C**) would have an impact on the overall tree topologies (**Supplementary Figure 1**), but there were no major differences in the topologies and only differences in support values for some branches. Species relationships in the broader Lamiaceae HPM assembly were rooted with *C. mutabilis* (tribe Ocimeae), followed by a clade which includes *Dracocephalum*, *Agastache*, *Lycopus*, and *Prunella* (1 local posterior probability [localPP] in every three), a second sister

clade with *Poliomintha* and *Hedeoma* (1 localPP), and the third clade with *Melissa* and *Lepechinia* (**Figure 2**). The four *Salvia* subgenera sampled (**Figures 2**, **3**; **Supplementary Figures 1 and 2**) are in "clade I" (clade nomenclature *sensu*; Walker et al., 2004; Jenks et al., 2013) with 1 localPP in every inference. Clade 1 included *S.* subg. *Salvia* (*S. officinalis*) and *Leonia* (*S. sessilifolia* and *S. texana*), sister to a clade of *S.* subg. *Audibertia* and *Calosphace* (1 localPP).

There were 8 out of the 13 *Salvia* subg. *Calosphace* sections *sensu* Epling which were sampled here and represented by more than one sample were monophyletic in all analyses (**Table 2**). They include *Cardinalis*, *Biflorae, Hastatae*, *Incarnatae, Lavanduloideae, Sigmoideae*, and *Uricae*, while *Curtiflorae* was only monophyletic in the nuclear ASTRAL and FastTree trees.

Several clades within *S.* subg. *Calosphace* was well-resolved and strongly supported by our phylogenetic results. A "*Hastatae* clade" with 1 PP (ASTRAL/BEAST) includes members of the *S.* sects. *Hastatae*, *Blakea,* and *Standleyana* are sisters to *S. axillaris* (monotypic *S.* sect. *Axillares*) (**Figure 2**). The "*Uliginosae* clade" includes a monophyletic *S.* sect. *Incarnatae* (*Salvia elegans* Vahl. + *Salvia cinnabarina* M. Martens and Galeotti) in all the analyses (1 localPP), and one sample each in *S.* sects. *Erythrostachys* (*Salvia regla* Cav.), *Cucullatae* (*Salvia clinopodioides* Kunth) and *Scorodoniae* (*Salvia ramosa* Brandegee). Following these *Calosphace* clades, we reached the "core *Calosphace*" (64 of the remaining species), where resolution and clade support are variable in the nuclear phylogenetic inferences (**Figures 2**, **3**; **Supplementary Figures 1, 2**).

Within the "core *Calosphace,*" the highly supported clades (1 localPP) included the "*Scorodoniae* clade" with species in *S.* sects. *Atratae*, *Mitratae,* and *Scorodoniae*. A large "*Fulgentes* clade" with monophyletic *S.* section *Cardinales* (with five of its nine species sampled) and some members of *S.* sects. *Fulgentes* and *Flocculosae* (1 local PP BEAST/ASTRAL). The "*Sigmoideae* clade" (1 local PP BEAST/ASTRAL) with *Salvia inconspicua* Benth. + *Salvia nepetoides* Kunth. and *Salvia aequidistans* Fernald (*S.* sect. *Scorodoniae*); a large clade with *Salvia gesneriiflora* Lindl. and Paxton in *S.* sect. *Nobiles* from Walker's "*Fulgentes* clade" [BEAST (0.881 local PP); ASTRAL (0.96 local PP)] and smaller strongly supported clades (1 local PP BEAST/ASTRAL) including monophyletic *S.* sects. *Uricae*, "*Lavanduloideae* clade," and "*Biflorae* clade," while *Curtiflorae* only in ASTRAL (0.96 localPP). Finally, the "*Polystachyae* clade" (1 localPP BEAST/ASTRAL), includes representatives from the *S.* sect. *Angulatae* (*S. tiliifolia*), *Iodanthae* (*Salvia iodantha* Fernald), *Polystachyae* (*S. brachyodonta*, *S. decora*, *Salvia filipes* Benth., *S. perblanda*, *Salvia plurispicata* Epling, *S. polystachia*, *S. purepecha*, *Salvia tonaticensis* Ramamoorthy ex Lara-Cabrera, Bedolla and Zamudio), and sect. *Purpureae* (*Salvia curviflora* Benth. and *S. purpurea*) and two samples each for *S. polystachia* (non-monophyletic) and *S. purpurea*, (monophyletic; **Figures 2**, **3**; **Supplementary Figure 1**).

The ML concatenated FastTree of the chloroplast loci (**Figure 3**) for the 90 samples, provided high support (1 localPP) for deep-level relationships within the *Ocimiae* and *Menthae*, and a sister relationship between *S.* subgenera *Audibertia* and *Calosphace*. Well-resolved and highly supported clades in this tree include *S. axillaris* as sister to the rest of subg. *Calosphace*; the "*Hastatae* clade" (1 localPP) and "*Uliginosae* clade" (1 localPP), with monophyletic *S.* sect. *Cardinales* (0.99 localPP), *Hastatae* (1 local PP), *Incarnatae* (1 local PP), and *Uricae* (1 localPP), and *S. hispanica* (two sampled). However, resolution and clade support are reduced for a few of the "core *Calosphace,*" such as the *S. genesneriflora* polytomy and *S. sect. Cucullatae* + *Scorodoniae*, *Flexulosae*, *Farinaceae*, *Albolanatae*. Two sections are not monophyletic for the cpDNA data *Lavanduloides* and *Sigmoideae*.

DISCUSSION

NGS in *Salvia*

The Hyb-Seq protocol (Weitemier et al., 2014) implemented, here, resolved deep phylogenetic relationships in Lamiaceae, among *Salvia* subgenera, and within a recently diverged *S.* subg. *Calosphace* (**Figure 2**; **Supplementary Figure 1**), providing additional support for existing phylogenetic hypotheses (Walker, 2006; Jenks et al., 2013; Fragoso-Martínez et al., 2018). We enriched 119 nuclear loci (**Supplementary Tables 3, 4**), 96 of which were left for phylogenetic estimations after the HPM filtering process.

To date, this is the largest base-pair sampling for this many *Salvia* species using the next-gen technology and specifically designed baits, and we were able to recover 1,314,368 bp (**Table 1**) in the HPM assembly for 96 nuclear genes in all 90 Lamiaceae sampled (14,604 b per sample; **Supplementary Table 4**). Previous anchored hybrid enrichment experiments in *Salvia* sampled 12 species for 453 loci producing a final alignment of 282,219 bp or 23,518 bp per sample (Fragoso-Martínez et al., 2017). Another study sampled 35 species (13 *Calosphace*) for 438 loci with a final alignment of 272,874 bp or 7,796 bp per sample (Kriebel et al., 2019). The studies by Fragoso-Martínez et al. (2017) and Kriebel et al. (2019) reported higher numbers of loci and base pairs than we did, but with less than half of our sampled taxa. Our methods had a more stringent cut-off for missing sequences and yielded a more conservative alignment. The branches in our tree with low support led to taxa that were not sampled in the study of Fragoso-Martinez or Kriebel et al. (2019).

We did not attempt a direct comparison between our custom-designed baits and previous next-gen studies using bait selection in Angiosperm v.1 kit (Buddenhagen et al., 2016). These three studies had different taxon sampling and phylogeny estimation methods so, it is not clear if the differences we report on branch support derive from our baits or taxon sampling.

Chloroplast Assembly

An additional advantage of the Hyb-Seq protocol as opposed to the AHE protocol, lies in obtaining the chloroplast and mitochondrial genomes, here we explored the chloroplast loci. Chloroplasts were assembled in HPM using *S. miltiorrhiza* genome as a pseudoreference, obtaining a 92,461 bp assembly for the 90 *Salvia* samples evaluated (**Supplementary Table 5**). A map to reference approach was previously tested (Olvera-Mendoza et al., 2020) on 15 samples from these same data to

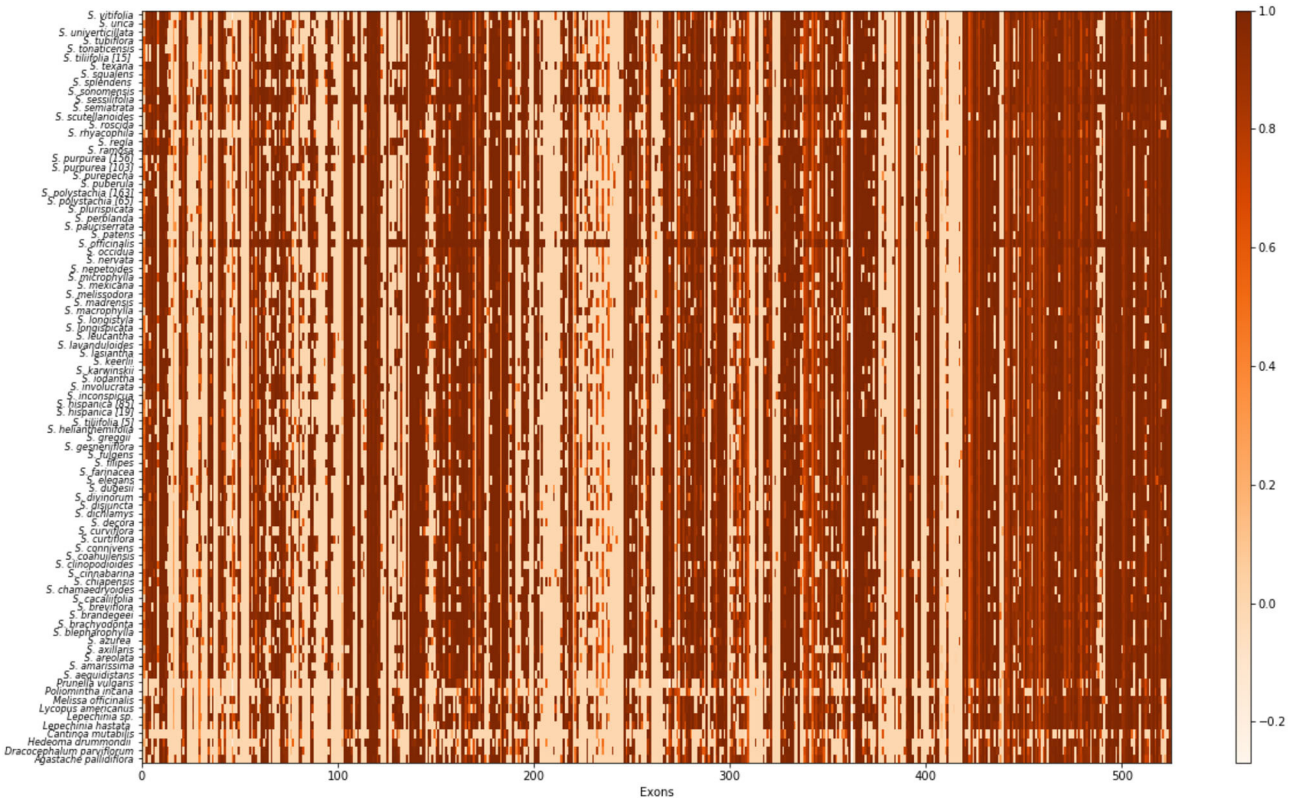


FIGURE 1 | Exon recovery heat map for 527 putative exons targeted by our baits. Each column represents an exon, and each row is each species. Each cell represents the sequence completeness; a lighter color signifies fewer bases recovered for that exon and a darker color signifies more bases were recovered.

investigate closely related species in *S.* sections *Atratae*, *Mitratae*, *Scorodoniae*, and *Sigmoideae*, resulting in the first chloroplast genome assemblies for *S.* subg. *Calosphace*, although limited taxon sampling for these sections impeded full resolution of the phylogeny. Our HPM chloroplast assembly using the same pseudoreference recovered fewer loci (**Supplementary Table 5**) than the study conducted by Olvera-Mendoza et al. (2020) did [114 genes, 80 CDS, 30 tRNA spacers, and 4rRNA's (Olvera-Mendoza et al., 2020) vs. our 75 CDS, 29 tRNA's, 5 introns and 4 rRNA]. This may be attributed to the many samples (78) we evaluated compared with their 15 samples, and the filtering step we used during HPM.

Nuclear Phylogenetic Inferences

The nuclear phylogenies (**Figure 2**; **Supplementary Figure 1**) resulting from ASTRAL and BEAST have well-resolved and highly supported clades and recover several previously reported relationships (Walker et al., 2004; Walker and Sytsma, 2007; Jenks et al., 2013; Will and Claßen-Bockhoff, 2017; Fragoso-Martínez et al., 2018). *Cantinoa* from tribe *Ocimeae* was used as the outgroup following Li et al. (2016). *Cantinoa* is sister to the *Mentheae* tribe and relationships in our trees are in agreement with the study of Drew and Sytsma (2012). We recovered subtribes Menthinae (*Hedeoma* and *Poliomintha*), Nepetinae (*Agastache* and *Dracocephalum*), Lycopinae (*Lycopus*), Prunellinae (*Prunella*), and Salviinae (*Melissa*, both *Lepechinia* and *Salvia*). Within *Salvia,* we recovered "clade I" with *S.* subgenera *Salvia* (*S. officinalis*) and *Leonia* (*S. sessilifolia* + *S. texana*), and a clade of *S.* subgenera *Audibertia* (*S. sonomensis* + *S. brandegeei*) and the 69 remaining species in *Calosphace*. Here we support the monophyly of eight of the 13 *Salvia* sections sampled (**Table 2**): *Biflorae*, *Curtiflorae*, *Hastatae, Incarnatae, Lavanduloideae, Sigmoideae,* and also *S.* sections *Cardinales* and *Uricae* (as in Olvera-Mendoza et al., 2020). Although our tree is well-resolved, our *Calosphace* sample is <15% of the estimated species diversity in the subgenus, undoubtedly having an effect on clade resolution, and unsampled species could potentially be inserted in future phylogenetic studies to further resolve fine-scale relationships with each clade.

Relationships among the section's sister to the core *Calosphace* have been somewhat controversial. Most studies (Walker et al., 2004; Walker and Sytsma, 2007; Jenks et al., 2013; Fragoso-Martínez et al., 2018; Kriebel et al., 2019) found *S. axillaris* (monotypic *S.* sect. *Axillares*) sister to the rest of the *Calosphace*; this relationship is only supported by our chloroplast analysis (**Figure 4**). Our nuclear data analyses (**Figure 2**; **Supplementary Figure 1**) support *S. axillaris* sister to "*Hastatae* clade," and together with sister to the rest of *Calosphace*; this relationship has also been recovered by Hu et al. (2018) [(*S. patens* + *Salvia cacaliifolia* Benth. (*S. axillaris* (rest of *Calosphace*)] and

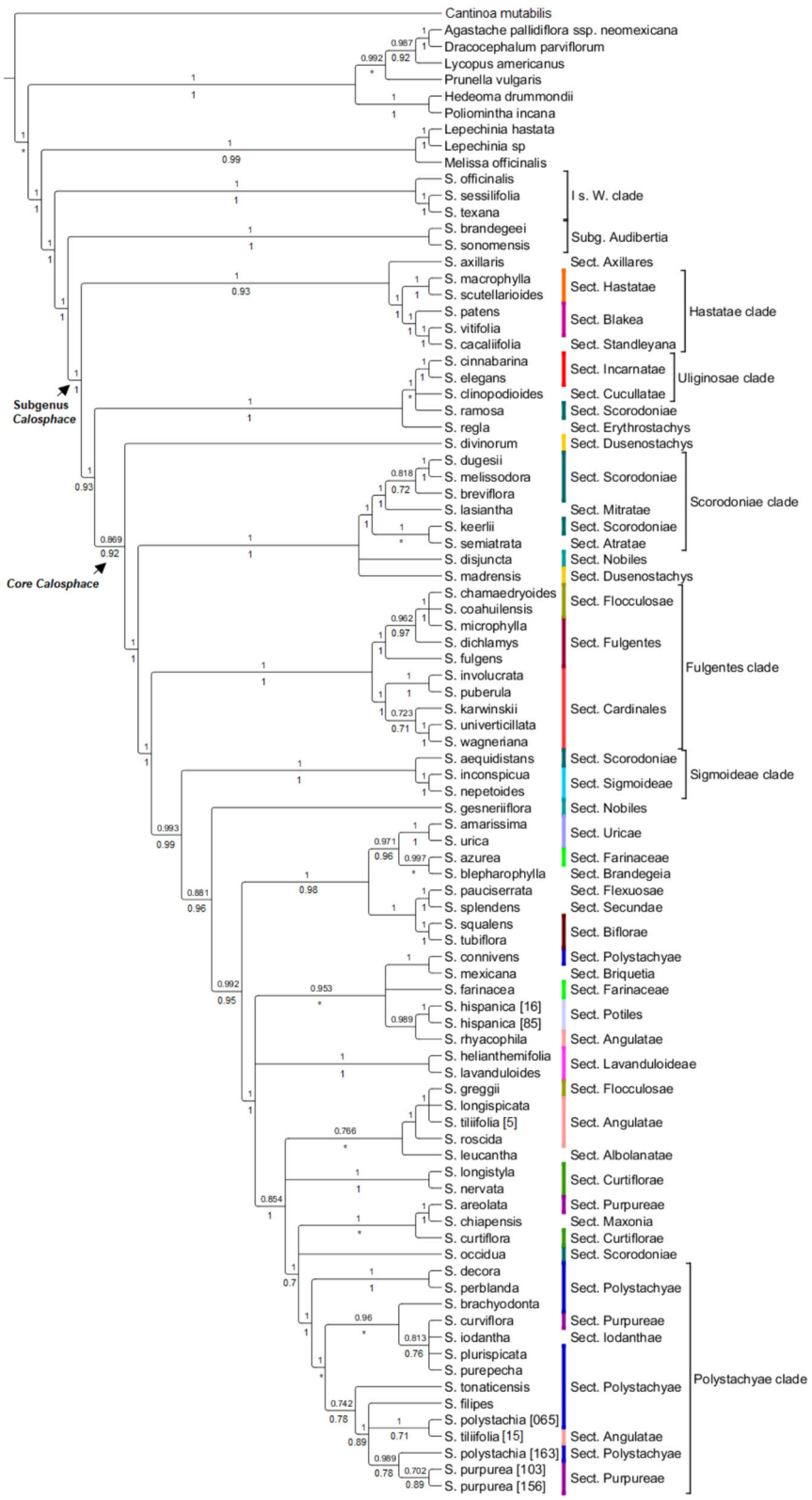


FIGURE 2 | HPM- BEAST tree for 96 nuclear genes with up to 70% missing data allowed for each exon. The BEAST local posterior probability is indicated above branches from the analysis and the ASTRAL analysis is under the branches. Branches with support values <0.7 are collapsed. *Salvia* subgenus *Calosphace* sections *s*. Epling are color-coded. The main clades follow previous nomenclature (Walker et al., 2004; Jenks et al., 2013; Fragoso-Martínez et al., 2018).

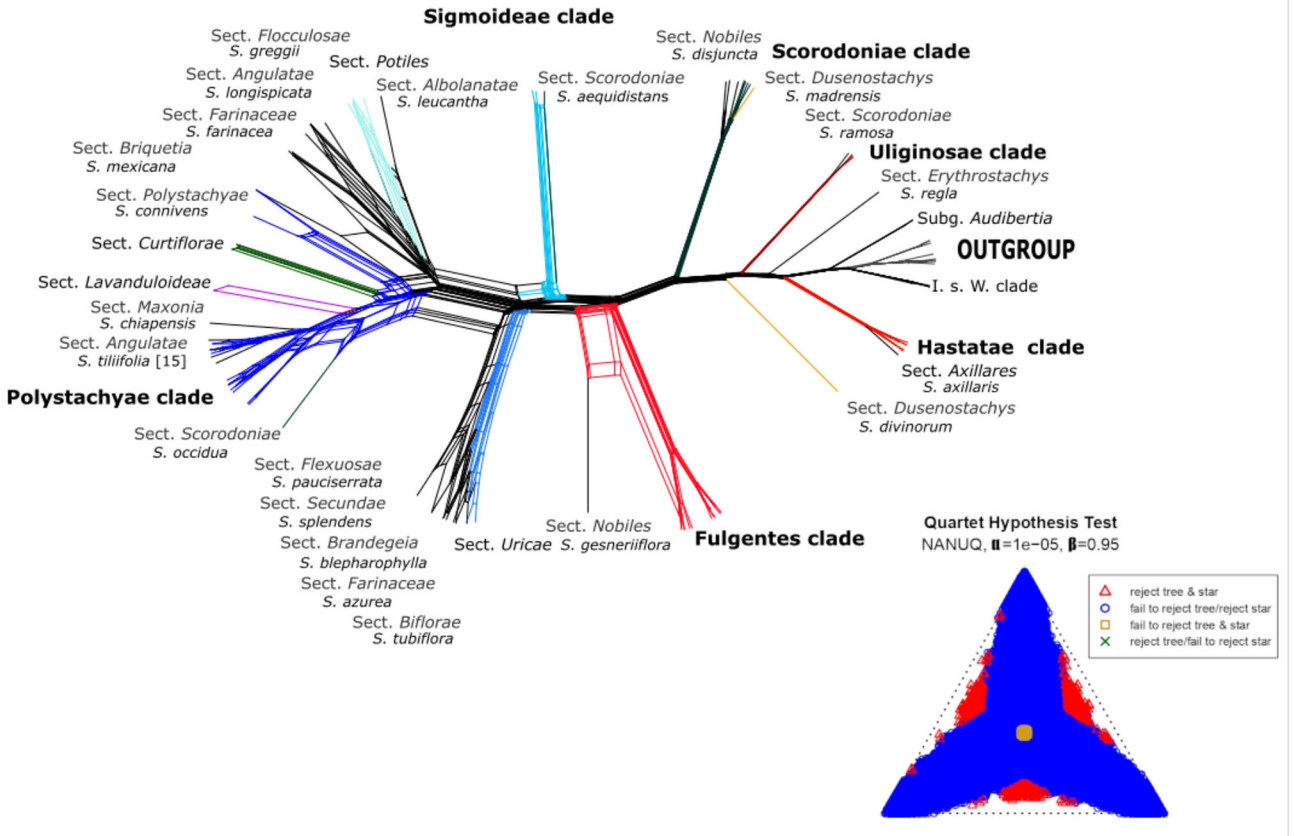


FIGURE 3 | Network analysis for 527 putative nuclear exons with *S.* subg. *Calosphace* sections are color-coded as in **Figure 1**. A simplex plot of NANUQ's quartet hypothesis test at $\alpha = 0.0001$ and $\beta = 0.95$.

Walker et al. (2015) [(*S. patens* (*S. axillaris* + *Salvia cedrosensis* Greene)] with *S. axillaris* in a clade with *Hastatae* representatives. Interestingly, these relationships are congruent with differences in stamen morphology; a key feature in *Salvia* (Bentham, 1832-1836; Fernald, 1900; Walker and Sytsma, 2007). Three stamen types have been described for *S.* subg. *Calosphace*; the G type in *S. axillaris* where both anterior and posterior anthers are expressed in free stamens, F type in the "clade Hastatae" (*S.* sects. *Standleyana*, *Blakea,* and *Hastatae*) where "both posterior thecae are aborted, and the adjacent posterior thecae are not, or only little fused" (Walker and Sytsma, 2007) and the E stamen type in the rest of the *Calosphace* where the posterior anthers are aborted and stamens are joined in a connective (Walker and Sytsma, 2007). The relationship we recovered suggests that elaborated connective tissue may have evolved twice in this clade (in *Hastatae* and *Calosphace*) or that the ancestor of the clade had another connective and it was lost in *S. axillaris*. The complex evolutionary patterns of stamen morphology are being investigated (Kriebel et al., 2020), to consider the potential usefulness of stamen characters for defining clades and within-species variation.

Previous next-gen studies of *Salvia* by Fragoso-Martínez et al. (2017) used the angiosperm bait kit (Johnson et al., 2019) and found three branches with low posterior probability (PP) support (their **Figure 1b**). Kriebel et al. (2019) on the other hand, found three poorly supported branches in the *Calosphace* clade in their ASTRAL coalescent analysis (**Figure 2**). We did not sample the taxa involved in two of those branches. Kriebel et al. (2019) additionally report an expanded taxon sampling to 266 *Calosphace* merging previous nuclear ribosomal DNA (ITS/ETS) sequences as supporting material for their habitat and pollinator study for *Salvia*.

Clade "*Hastatae*" was recovered in every tree (**Figures 2**, **4**; **Supplementary Figures 1–3**) and includes reciprocally monophyletic *S.* sects. *Hastatae*, *Blakeae,* and *Standleyana*. This clade was also found in other studies [Jenks et al., 2013; Fragoso-Martínez et al., 2018; Kriebel et al., 2019 (nrDNA)]. *Salvia* sect. *Standleyana* was redefined by Turner (2011), merging it with species from the *S.* sect. *Blakea s*. Epling (*Salvia costaricensis* Oerst., *S. patens*, *S. subpatens* Epling, *S. vitifolia*). Later Klitgaard (2012) supported the merger of these *S.* sections, but under sect. *Blakea*. Our phylogenies found *S. cacalifolia* in a clade with *S. patens* and *S. vitifolia* and so support the merger of sect. *Standleyana* and *Blakea*, with the caveat, that *S. costaricensis* Oerst., *S. subpatens*, and *S. serboana* B. L. Turner should be sampled in a molecular study before the sections are re-classified.

Our clade "*Uliginosae*" (**Figures 2**, **4**; **Supplementary Figures 1–3**) includes monophyletic *S.* sects. *Incarnatae* and *Cucullatae*,

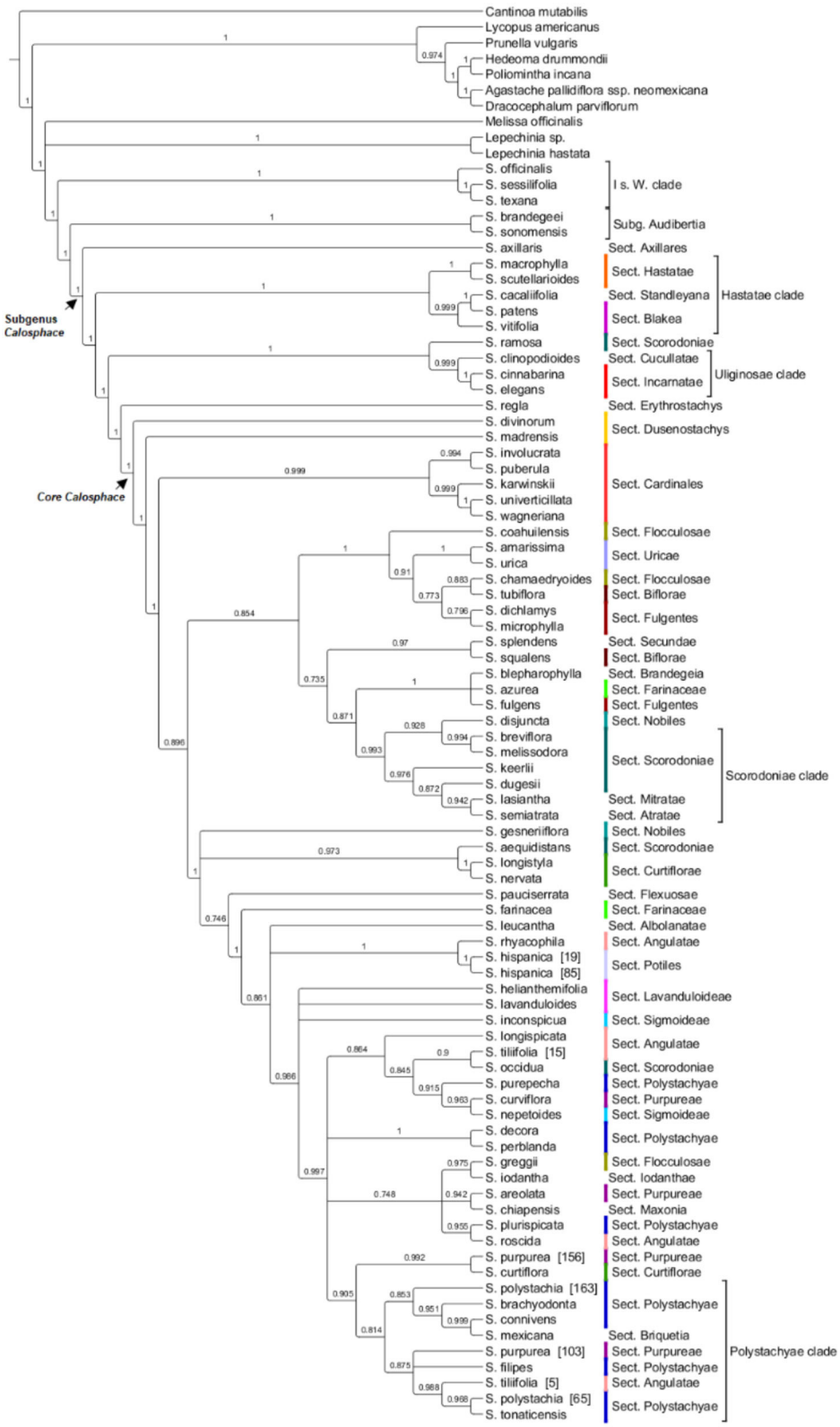


FIGURE 4 | HPM- ML FastTree for 114 chloroplast loci, local posterior probability >0.7 is indicated above branches (lower are collapsed). *Salvia* subgenus *Calosphace* sections *s*. Epling are color-coded. The main clades follow previous nomenclature (Walker et al., 2004; Jenks et al., 2013; Fragoso-Martínez et al., 2018).

TABLE 2 | *Salvia* subgenus *Calosphace* monophyletic sections *s*. Epling, comparative of previous phylogenetic analysis and our Hyb-seq three nuclear and chloroplast analyses.

***Salvia* sect. *sensu* Epling**	**Fragoso-Martínez et al. (2018)**	**nASTRAL**	**nBEAST**	**cpFastTree**
Angulatae (4/52)	No	No	No	No
Biflorae (2/4)	Yes	Yes	Yes	No
Blakea (2/5)	Yes	No	No	No
Cardinalis (5/9)	No	Yes	Yes	Yes
Curtiflorae (3/9)	Yes	Yes	No	No
Fulgentes (3/9)	No	No	No	No
Hastatae (2/7)	Yes	Yes	Yes	Yes
Incarnatae (2/2)	Yes	Yes	Yes	Yes
Lavanduloideae (2/18)	Yes	Yes	Yes	Yes
Polystachyae (9/16)	No	No	No	No
Purpureae (3/9)	No	No	No	No
Sigmoideae (2/9)	Yes	Yes	Yes	Yes
Uricae (2/2)	No	Yes	Yes	Yes

agreeing with previous clade circumscription [Jenks et al., 2013; Fragoso-Martínez et al., 2018; Kriebel et al., 2019 (nrDNA)]; unfortunately, though, our sampling in this clade is reduced, and we are lacking a representative of *S.* sect. *Uliginosae*; furthermore, our trees include *S. regla* and one of the seven sampled *S.* sect. *Scorodoniae* (*S. ramosa*) in the *S.* sect. *Erythrostachys* clade; these relationships require careful review with broader taxon sampling within the *S.* sect. *Erytrostachys*.

Following clades "*Hastate*" and "*Uliginosae*" we reach the troublesome and most species-rich clade, the "core *Calosphace*" (**Figures 2–4**; **Supplementary Figures 1–3**). The remainder of the sampled species is included in this clade. Walker (2006) was the first to define this clade, consisting of several clades immersed within a large polytomy and later studies with expanded sampling have confirmed this clade (Jenks et al., 2013; Fragoso-Martínez et al., 2018). We newly placed seven species in the *Calosphace* clade, classified in the *S.* sects. *Angulatae* (*S. roscida*), *Cardinales* (*S. puberula*), *Fulgentes* (*S. dichlamys*) and *Polystachyae* (*S. brachyodonta*, *S. decora*, *S. perblanda*, *S. purepecha*). Additionally, we found monophyletic *S.* sects. *Cardinalis* and *Uricae*, increasing the molecular evidence for monophyletic *Calosphace* sections from 12 (Fragoso-Martínez et al., 2018) to 14 among those evaluated.

A close sectional relationship has been demonstrated for *Salvia* sects. *Scorodoniae Atratae* (*S. semiatrata*), *Mitratae* (*Salvia lasiantha* Benth.), *Sigmoideae* (*S. inconspicua* and *S. nepetoides*), and *Uricae* (*S. amarissima* and *S. urica*) cpDNA entire genome and nuclear ribosomal cistron (Olvera-Mendoza et al., 2020). We found support for relationships among some of these sections, but together they do not form a clade; *S.* sect. *Uricae* is indeed monophyletic and distinct from the *S.* sect. *Scorodoniae* as Olvera-Mendoza et al. (2020) proposed. *Salvia* sect. *Scorodoniae* is not monophyletic although morphologically recognizable (Olvera-Mendoza et al., 2017) and *S.* sect. *Sigmoideae* is monophyletic only if nuclear data are incorporated in the analysis, either combined cpDNA + nDNA (Jenks et al., 2013; Fragoso-Martínez et al., 2018; Olvera-Mendoza et al., 2020) or only nuclear (**Figure 2**; **Supplementary Figures 1–3A–C**); highlighting the importance of nuclear markers to better-resolve *Salvia* species relationships. Jenks et al. (2013) and Fragoso-Martínez et al. (2018) also recovered a non-monophyletic *S.* sect. *Scorodoniae* [as did Kriebel et al., 2019 (nrDNA)] and considered *S.* sect. *Uricae*'s species to be best placed within *S.* sect. *Scorodoniae*. It is clear that further analysis is required to solve species relationships within these sections, strive to fully sample *S.* sects *Scorodoniae* and *Sigmoideae*, coupled with a thorough morphological review.

Our topology for "*Fulgentes* clade" (**Figure 2**; **Supplementary Figure 1**) is similar to previous inferences but with high branch support (Jenks et al., 2013; Fragoso-Martínez et al., 2018; Kriebel et al., 2019 [nrDNA]) for the nuclear loci analyses, including members in *S.* sects. *Fulgentes*, *Flocculosae,* and *Cardinalis* (their *Holwaya s.* Ramamoorthy, 1984). Fragoso-Martínez et al. (2017) AHE analysis report *S. fulgentes* sister to the rest of core *Calosphace* except *S. melissodora* and *S. mocinoi* in their branch B3 (0.71 PP). *Salvia* sect. *Cardinales* is here represented by five (*Salvia involucrata* Cav., *Salvia karwinskii* Benth., *S. puberula*, *Salvia wagneriana* Pol., *Salvia univerticillata* Ramamoorthy ex Klitg.) of its nine species and is monophyletic and strongly supported in all nuclear (**Figure 2**; **Supplementary Figures 1, 2A–C**) and chloroplast trees (**Figure 4**; **Supplementary Figures 3A–C**). Section *Cardinales* is sister to a clade of *S.* sects. *Fulgentes* (*Salvia fulgens* Cav., *S. dichlamys*, *Salvia microphylla* Kunth) and *Flocculosae* (*Salvia chamaedryoides* Cav., *Salvia coahuilensis* Fernald), only our *Salvia greggii* A. Gray (*S.* sect. *Flocculosae*) is apart from this clade. Despite the non-monophyly of *S.* sects. *Flocculosae* and *Fulgentes* we agree with Jenks et al. (2013) on their morphological and phylogenetic relationships.

One of the most species-rich sections in *Salvia* subg. *Calosphace* is *Angulatae* (52 species) and it is also one of the most morphologically complex and has a disjunct distribution in N and S America (Epling, 1939; Walker, 2006). None of the previous studies have recovered it as monophyletic

[Walker, 2006; Jenks et al., 2013; Fragoso-Martínez et al., 2018; Kriebel et al., 2019 (nrDNA)]. Here we found three species, *S. roscida, S. longispicata and S. tiliifolia* [5] form the broadly defined "*Angulatae* clade" (**Figure 2**; **Supplementary Figure 1**) and *S. tiliifolia* [15] is sister to *S. polystachia* [163] within the "*Polystachyae* clade." The non-monophyly of *S. tiliifolia* is both troublesome and expected since Walker (2006) found a monophyletic *S. tiliifolia* lacking bootstrap support in his neighbor-joining tree, and *S. tiliifolia* is one of the most broadly distributed and morphologically complex species in subg. *Calosphace*. Section *Angulatae* is in urgent need of a thorough review, both morphologically and molecularly; to date, only 22 South American members have been studied (Fernández-Alonso, 2003; Wood, 2007) and there are ∼26 North American members that remain to be sampled.

Finally, the "*Polystachyae* clade" (**Figures 2–4**; **Supplementary Figures 1–3**) includes members from *S.* sects. *Angulatae* (*S. tiliifolia* [15]), *Curtiflorae* (*S. curtiflora*), *Iodanthae* (*S. iodantha*), *Maxonia* (*Salvia chiapensis* Brandegee), *Purpureae* (*S. curviflora*, *S. purpurea*), and *Scorodoniae* (*S. occidua*). Three of these sections have been under study for some time since Walker (2006) first found *S. iodantha*, *S. polystachia,* and *S. purpurea* in a clade with only 1-2 bp difference in *psbA-trnH, trnL-trnF*, and ITS sequences. Later Bedolla-García (2012) expanded taxon sampling and regarded this as the "PIP clade," due to the inclusion of members of *S.* sects. *Purpureae* from Mexico (*S. areolata*, S. *curviflora*, *S. littae*, *S. purpurea*, *S. raveniana*), *Iodanthae* (*S. iodantha*, considering *Salvia arbuscula* Fernald and *Salvia townsendii* Fernald as synonyms) and *Polystachyae* (*S. brachyodonta*, *Salvia connivens* Epling, *Salvia compacta* Kuntze, *S. decora*, *S. filipes*, *Salvia mcvaughii* Bedolla, Lara Cabrera and Zamudio, *S. plurispicata*, *S. polystachia*, *Salvia tonalensis* Brandegee, *S. tonaticensis*). Here we include nine of the sixteen species in the *S.* sect. *Polystachyae*, three species of *S.* sect. *Purpureae* and *S. iodantha* (sole species in *S.* sect. *Iodanthae*), and all sampled taxa of these sections, with the exception of *S. connivens* (*S.* sect. *Polystachyae*), are in this clade. Neither *S.* sects. *Purpureae* nor *Polystachyae* are monophyletic, as has been the case elsewhere [Jenks et al., 2013; Fragoso-Martínez et al., 2018; Kriebel et al., 2019 (nrDNA)]. For this troublesome, widely diverse clade we recovered only one consistent and supported sister relationship (*S. decora* and *S. perblanda*) in the nuclear trees (**Figures 2**, **3**; **Supplementary Figures 1, 2**), network (**Figure 3**), and also in the cpDNA tree (**Figure 4**; **Supplementary Figure 3**). Otherwise, species relationships in this part of the tree have less support, with some polytomies and low to medium branch support (**Figures 2**, **3**; **Supplementary Figures 1, 2**). This lack of branch support and the network results strongly suggest reticulation issues due to recent divergence, hybridization, or incomplete lineage sorting (Huang et al., 2017). Additionally, we found that *S. purpurea* is monophyletic in the nuclear evidence, whereas *S. polystachia* is not.

Aside from the main clades "*Hastatae*," "*Uliginosae*," "*Scorodoniae*," "*Fulgentes*," "*Sigmoideae*," and "*Polystachyae*" we found other strongly supported, small clades. *Salvia* sect. *Uricae* is monophyletic and *S.* sects. *Farinaceae*, *Nobiles*, and *Dusenostachys* are non-monophyletic, as has been previously reported [Jenks et al., 2013; Fragoso-Martínez et al., 2018; Kriebel et al., 2019 (nrDNA)]. We also support the monophyly of *Salvia hispanica* (*S.* sect. *Potiles*), the two samples forming a clade with *S. rhyacophila* (*S.* sect. *Angulatae*) as did Fragoso-Martínez et al. (2018); whereas Fragoso-Martínez et al. (2017) AHE analysis found a poorly supported sister relationship between *S. hispanica* and *S. heliamenthifolia* (0.53).

Chloroplast Phylogeny

Following Doyle (2021) we opted to analyze our chloroplast data as a single hereditary unit through ML in FastTree (**Figure 3**). The chloroplast tree supports the outgroup relationships *S. axillaris* as sister to the rest of *S.* subg. *Calosphace* and sister lineages and clades "*Hastatae*" and "*Uliginosae*," and monophyletic *S.* sects. *Cardinales*, *Hastatae*, *Incarnatae* and *Uricae*. Only two *S.* sects. are not monophyletic here as opposed to nDNA, *Lavanduloides*, and *Sigmoideae*. Our nuclear and chloroplast analyses, however, used distinct pseudo references, here, we used the distantly-related *S. miltiorrhiza* (*Salvia* subg. *Sclarea* sect. *Drymosphace* Hu et al., 2018) as the chloroplast assembly pseudoreference. *Salvia miltiorrhiza* is sister to clade *Meriandra* + *Dorystaechas* + *Ramona* (*Salvia* subg. *Audibertia*) + *Lasemia* (*Salvia* subg. *Calosphace*) (Will and Claßen-Bockhoff, 2017).

Our nuclear and chloroplast phylogenies are in overall agreement, for the outgroup, sister relationship of *Audibertia* and *Calosphace* and well-resolved "*Hastatae*," "*Uliginosae*," "*Scorodoniae,*" and "*Polystachyae*" clades. However, they disagree on the placement of *S. axillaris* as sister to "clade *Hastatae*" in nuclear trees or sister to the rest of the *Calosplace* in the chloroplast tree. Within the core *Calosphace*, particular complexity in the phylogenies and network is seen with *Salvia gesneriiflora*, a bird pollinated and morphologically distinct species. This species is one of the two representatives of the *S.* sect. *Nobiles* in our sampling (*S. disjuncta* is the other) and S. *gesneriiflora* placement moves between the "*Sigmoideae*" and "*Uricae* clades" in BEAST (**Figure 2**), between the "*Fulgentes* clade" and "*Sigmoideae* clade" in ASTRAL (**Supplementary Figure 1**), and between the "*Scorodoniae* clade" and *Scorodoniae*+*Curtiflorae* clade in the chloroplast tree (**Figure 4**). Furthermore, the network shows the nuclear loci for this species have characters that align it with *S. coahuilensis* in clade *Flocculosae* + *Uricae* + *Fulgentes* and also align it with the remaining core *Calosphace* clade (**Figure 3**). It is unclear why the placement of this particular species is so troublesome, no hybridization events have been reported, though frequent nectar robbing does occur (Cuevas et al., 2013), so hybridization may be a possibility worth further exploration. It is possible that we lacked sampling of phylogenetically closer relatives. Interestingly, the sectional circumscription of this species has also been controversial, Santos (1991) moved *S. gesneriiflora* from the *S.* sect. *Nobiles* Epling (1939) to sect. *Holwayana*. Testing the placement of this species would require a phylogeographic approach.

Species Monophyly

This research addressed *Salvia* taxon monophyly with NGS data. Within *Calosphace* monophyly has been an issue for *S.* sections *sensu* Epling and species, particularly in sections with disjunct

distribution and widely distributed and variable species. The discordance between morphological recognition of sections *s.* Epling and later molecular phylogenies have also been discussed elsewhere (Jenks et al., 2013; Fragoso-Martínez et al., 2018) and has been hypothesized to be caused by morphological homoplasy due to pollinator pressure.

Species monophyly has been addressed several times in *S.* subg. *Calosphace* through traditional Sanger sequencing, mostly rejecting monophyly. For example, Walker (2006) sampled several specimens each of *S. polystachia, S. purpurea,* and *S. tiliifolia*, and only the latter was monophyletic in his neighbor-joining tree. Later Jenks et al. (2013) found *S. microphylla*, *S. mexicana,* and *S. polystachia* to be non-monophyletic. In our results, *S. hispanica* and *S. purpurea* are monophyletic whereas traditional Sanger (Walker, 2006) sequencing rejected *S. purpurea* monophyly. However, our massive alignment was not sufficient to test monophyly for *S. polystachia* nor *S. tiliifolia*. Species monophyly for these and other species will likely need a distinct approach, such as phylogeography (Cutter, 2013), to get a better grasp at the speciation processes, particularly for such morphologically complex and amply distributed species.

In this study, we provide valuable new evidence as to the utility of Hyb-Seq data for capturing 96 nuclear loci from phylogenetically distant Lamiaceae and closely related *Salvia* subg. *Calosphace*, including testing species monophyly. We also recovered the cpDNA genome with concatenated tree phylogeny in agreement with the nuclear genome with this sampling and with previous phylogenies and improved clade resolution. We found two newly supported monophyletic *S.* subg. *Calosphace* sections *s*. Epling and two of four species tested were monophyletic. Although this is the largest NGS study of *Salvia* to date, a more thorough taxon sampling is necessary to better test sectional relationships. NGS-based approaches combined with the reassessment of morphological characters are needed to re-assess sectional circumscription, study the complex species groups in subg. *Calosphace*, and eventually produce a new monograph. Beyond the implications for systematics, a robust phylogeny for the genus is necessary to test hypotheses about the evolution of pollinator associations and morphological adaptations to pollinators. We hope that sage researchers will use our bait design across the width of the phylogenetic spectrum as a steppingstone to build upon for future studies.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: BioSample: PRJNA748827.

AUTHOR CONTRIBUTIONS

GG, JP, and SL-C conceived the study and data analysis. SL-C contributed to the laboratory work. CM-L and MP-G contributed to data analysis. AF, AC-J, and JM-C contributed to analysis and manuscript review. All authors contributed to manuscript writing and review, read, and approved the final manuscript.

FUNDING

This work was supported by the following: CONACyT graduate studies scholarship 618610 MP-G, CONACyT sabbatical scholarship 232839 to SL-C, Coordinación de la Investigación Científica (UMSNH), Project 8.16. NSF Award Number 1120080 to JP. And NSF Grant DEB-1557059 that supports CM-L post-doctoral position.

ACKNOWLEDGMENTS

Some of the results here presented are part of the M.S. thesis of LPG under the advisory of SL-C. We appreciate the help of curators from herbaria RSA, BIGU, EBUM, ENCB, IEB, MEXU, and UAMIZ. We are thankful for the field collection assistance of Arnulfo Blanco and Botanic Garden collection to Heather Blume for facilitating collections at the Cabrillo College Environmental Horticulture Center and Botanic Gardens, Holly Forbes from Berkeley Botanic Garden (UCB), and disposition for collecting at Huntington Botanic Garden and RSA.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.725900/full#supplementary-material

REFERENCES

Allman, E. S., Baños, H., and Rhodes, J. A. (2019). NANUQ: a method for inferring species networks from gene trees under the coalescent model. *Algorithms Mol. Biol.* 14:24. doi: 10.1186/s13015-019-0159-2

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed October 12, 2020).

APG (1998). An ordinal classification for the families of flowering plants. *Ann. Mo. Bot. Gard.* 85, 531–553. doi: 10.2307/2992015

Bedolla-García, B. Y. (2012). *Filogenia de Salvia secc. Polystachyae (Lamiaceae)* (Tesis de doctorado). Facultad de biología, Universidad Michoacana de San Nicolás de Hidalgo. Morelia, Michoacán, México, 1–165.

Benítez-Vieyra, S., Fornoni, J., Pérez-Alquicira, J., Boege, K., and Domínguez, C. A. (2014). The evolution of signal-reward correlations in bee-and hummingbird-pollinated species of *Salvia*. *Proc. Biol. Sci.* 281:20132934. doi: 10.1098/rspb.2013.2934

Bentham, G. (1832-1836). "*Salvia*," in *Labiat. Gen. J*, eds G. Bentham (Ridgway and Sons Picadilli, London), p. 190–312.

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

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*. 28:e1660. doi: 10.7287/peerj.preprints.1355

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10:e1003537. doi: 10.1371/journal.pcbi.1003537

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Breinholt, J. W., Carey, S. B., Tiley, G. P., Davis, E. C., Endara, L., McDaniel, S. F., et al. (2021). A target enrichment probe set for resolving the flagellate land plant tree of life. *APPS* 9:e11406. doi: 10.1101/2020.05.29.124081

Buddenhagen, C., Lemmon, A. R., Lemmon, E. M., Bruhl, J., Cappa, J., Clement, W. L., et al. (2016). Anchored phylogenomics of angiosperms I: assessing the robustness of phylogenetic estimates. *bioRxiv* 1–65. doi: 10.1101/086298

Carlsen, M. M., Fér, T., Schmickl, R., Leong-Škorničková, J, Newman, M, and Kress, W. J. (2018). Resolving the rapid plant radiation of early diverging lineages in the tropical Zingiberales: pushing the limits of genomic data. *Mol. Phylogenet. Evol.* 128, 55–68. doi: 10.1016/j.ympev.2018.07.020

Carter, K. A., Liston, A., Bassil, N. V., Alice, L. A., Bushakra, J. M., Sutherland, B. L., et al. (2019). Target capture sequencing unravels Rubus evolution. *Front. Plant Sci.* 10:1615. doi: 10.3389/fpls.2019.01615

Celep, F., Atalay, Z., Dikmen, F., Dogan, M., Sytsma, K. J., and Claßen-Bockhoff, R. (2020). Pollination ecology, specialization, and genetic isolation in sympatric bee-pollinated *Salvia* (Lamiaceae). *Int. J. Plant Sci.* 181, 800–811. doi: 10.1086/710238

Chamala, S., García, N., Godden, G. T., Krishnakumar, V., Jordon-Thaden, I. E., De Smet, R., et al. (2015). MarkerMiner 1.0: a new application for phylogenetic marker development using angiosperm transcriptomes. *APPS* 3:1400115. doi: 10.3732/apps.1400115

Claßen-Bockhoff, R., Speck, T., Tweraser, E., Wester, P., Thimm, S., and Reith, M. (2004). The staminal lever mechanism in *Salvia* L. (Lamiaceae): a key innovation for adaptive radiation? *Org. Divers. Evol.* 4, 189–205. doi: 10.1016/j.ode.2004.01.004

Constantinides, B., and Robertson, D. L. (2017). Kindel: indel-aware consensus for nucleotide sequence alignments. *JOSS* 2:282. doi: 10.21105/joss.00282

Couvreur, T. L. P., Helmstetter, A. J., Koenen, E. J. M., Bethune, K., Brandão, R. D., and Little, S. A. (2019). Phylogenomics of the major tropical plant family annonaceae using targeted enrichment of nuclear genes. *Front. Plant Sci.* 9:1941. doi: 10.3389/fpls.2018.01941

Cuevas, G. E., Alcalá-Guerra, A., Baños-Bravo, Y., and Flores, P. A. (2013). Biología reproductiva y robo de néctar en *Salvia gesneriflora* (Lamiaceae) y sus consecuencias en el éxito reproductivo. *Bot. Sci*., 91, 357–362. doi: 10.17129/botsci.14

Cutter, A. D. (2013). Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. *Mol. Phyl. Evol.* 69, 1172–1185. doi: 10.1016/j.ympev.2013.06.006

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods*. 9:772. doi: 10.1038/nmeth.2109

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A*. 110, 898–903. doi: 10.1073/pnas.1300127110

Doyle, J., and Doyle,. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* 19, 11–15.

Doyle, J. J. (2021). Defining coalescent genes: theory meets practice in organelle phylogenomics. *Syst. Biol.* syab053. doi: 10.1093/sysbio/syab053

Drew, B. T., González-Gallegos, J. G., Xiang, C. L., Kriebel, R., Drummond, C. P., Walker, J., et al. (2017). *Salvia* united: the greatest good for the greatest number. *Taxon* 66, 133–145. doi: 10.12705/661.7

Drew, B. T., and Sytsma, K. J. (2012). Phylogenetics, biogeography, and staminal evolution in the tribe *Mentheae* (Lamiaceae). *Am. J. Bot.* 99, 933–953. doi: 10.3732/ajb.1100549

Drummond, A. J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214. doi: 10.1186/1471-2148-7-214

Epling, C. (1939). A revision of *Salvia*, subgenus Calosphace. *Beihefte Feddes Repertorium Specier. Novarum regni Veg.* 110, 1–383.

Epling, C. (1940). Supplementary notes on American Labiatae. *Bull. Torrey Bot. Club.* 67, 509–534. doi: 10.2307/2480972

Epling, C. (1941). Supplementary notes on American Labiatae-II. *Bull. Torrey Bot. Club.* 68, 552–568. doi: 10.2307/2481456

Epling, C. (1944). Supplementary notes on American Labiatae-III. *Bull. Torrey Bot. Club.* 71, 484–497. doi: 10.2307/2481241

Epling, C. (1947). Supplementary notes on American Labiatae-IV. *Bull. Torrey Bot. Club.* 74, 512–518. doi: 10.2307/2481876

Epling, C. (1951). Supplementary notes on American Labiatae-V. *Brittonia* 7, 129–142. doi: 10.2307/2804702

Epling, C., and Jativa, M. C. (1963). Supplementary notes on American Labiatae-VIII. *Brittonia* 15, 366–376. doi: 10.2307/2805381

Epling, C., and Mathias, M. E. (1957). Supplementary notes on American Labiatae-VI. *Brittonia* 8, 297–313. doi: 10.2307/2804980

Fér, T., and Schmickl, R. E. (2018). Hybphylomaker: target enrichment data analysis from raw reads to species trees. *Evol. Bioinform.* 14, 1–9. doi: 10.1177/1176934317742613

Fernald, M. L. (1900). A synopsis of the Mexican and Central American species of *Salvia*. *Proc. Am. Acad. Arts Sci.* 35, 489–556. doi: 10.2307/25129966

Fernández-Alonso, J. L. (2003). Estudios en Labiatae de Colombia IV. Novedades en *Salvia* y Sinópsis de las secciones Angulatae y Purpurea. *Caldasia* 25, 235–281.

Fisher, A. E., Hasenstab, K. M., Bell, H. L., Blaine, E., Ingram, A. L., and Columbus, J. T. (2016). Evolutionary history of chloridoid grasses estimated from 122 nuclear loci. *Mol. Phylogenet. Evol.* 105, 1–14. doi: 10.1016/j.ympev.2016.08.011

Fragoso-Martínez, I., Martínez-Gordillo, M., Salazar, G. A., Sazatornil, F., Jenks, A. A., García-Peña, M. R., et al. (2018). Phylogeny of the Neotropical sages (*Salvia* subg. Calosphace; Lamiaceae) and insights into pollinator and area shifts. *Plant. Syst. Evol.* 304, 1–13. doi: 10.1007/s00606-017-1445-4

Fragoso-Martínez, I., Salazar, G. A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Lemmon, E. M., et al. (2017). A pilot study applying the plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus Calosphace; Lamiaceae). *Plant. Syst. Evol.* 117, 124–134. doi: 10.1016/j.ympev.2017.02.006

Frodin, D. G. (2004). History and concepts of big plant genera. *Taxon* 53, 753–776. doi: 10.2307/4135449

Ge, X., Chen, H., Wang, H., Shi, A., and Liu, K. (2014). *De novo* assembly and annotation of *Salvia splendens* transcriptome using the illumina platform. *PLoS ONE* 9:e0087693. doi: 10.1371/journal.pone.0087693

González-Gallegos, J. G., Bedolla-García, B. Y., Cornejo-Tenorio, G., Fernández-Alonso, J. L., Fragoso-Martínez, I., García-Peña, M. R., et al. (in press). Richness and distribution of *Salvia* subgenus Calosphace (Lamiaceae). *Int. J. Plant Sci.*. 181:831–856. doi: 10.1086/709133

Harley, R. M., Atkins, S., Budantsev, A. L., Cantino, P. D., Conn, B. J., Grayer, R., et al. (2004). "Labiatae," in *Labiatae. The Families and Genera of Vascular Plants VII. Flowering Plants Dicotyledons: Lamiales (Except Acanthaceae Including Avicenniaceae)*, eds. K. Kubitzkiy, J. W. Kadereit. (Berlin: Springer). p. 167–4275.

Herrando-Moraira, S and The Cardueae Radiations Group (2018). Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae). *Mol. Phylogenet. Evol.* 128, 69–87. doi: 10.1016/j.ympev.2018.07.012

Herrando-Moraira, S and The Cardueae Radiations Group (2019). Nuclear and plastid DNA phylogeny of tribe Cardueae (Compositae) with Hyb-Seq data: a new subtribal classification and a temporal diversification framework. *Mol. Phylogenet. Evol.* 137, 313–332. doi: 10.1016/j.ympev.2019.05.001

Hu, G. X., Takano, A., Drew, B. T., Liu, E.-D., Soltis, D. E., Soltis, P. S., et al. (2018). Phylogeny and staminal evolution of *Salvia* (Lamiaceae, Nepetoideae) in East Asia. *Ann. Bot.* 122, 649–668. doi: 10.1093/aob/mcy104

Huang, H., Sukumaran, J., Smith, S. A., and Knowles, L. L. (2017). Cause of gene tree discord? Distinguishing incomplete lineage sorting and lateral gene transfer in phylogenetics. *PeerJ*. e3489v1. doi: 10.7287/peerj.preprints.3489

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Jenks, A. A., Walker, J. B., and Kim, S. C. (2013). Phylogeny of new world *Salvia* subgenus Calosphace (Lamiaceae) based on cpDNA (psbA-trnH) and nrDNA (ITS) sequence data. *J. Plant Res.* 126, 483–496. doi: 10.1007/s10265-012-0543-1

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., et al. (2019). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* 68, 594–606. doi: 10.1093/sysbio/syy086

Jones, K. E., Fér, T., Schmickl, R. E., Dikow, R. B., Funk, V. A., Herrando-Moraira, S., et al. (2019). An empirical assessment of a single family-wide hybrid capture locus set at multiple evolutionary timescales in Asteraceae. *Appl. Plant Sci.* 7:e11295. doi: 10.1002/aps3.11295

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 15:3059-66. doi: 10.1093/nar/gkf436

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kent, W. J. (2002). BLAT - The BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Klitgaard, B. (2012). "*Salvia* L," in: *Flora Mesoamericana*, eds. G. Davidse, S. M. Sousa, S. Knapp, F. Chiang (Rubiaceae a Verbenaceae. Missouri Botanical Press., St. Louis), p. 396–424.

Kriebel, R., Drew, B. T., Drummond, C. P., González-Gallegos, J. G., Celep, F., Mahdjoub, M. M., et al. (2019). Tracking temporal shifts in area, biomes, and pollinators in the radiation of *Salvia* (sages) across continents: leveraging anchored hybrid enrichment and targeted sequence data. *Am. J. Bot.* 106, 573–597. doi: 10.1002/ajb2.1268

Kriebel, R., Drew, B. T., González-Gallegos, J. G., Celep, F., Heeg, L., Mahdjoub, M. M., et al. (2020). Pollinator shifts, contingent evolution, and evolutionary constraint drive floral disparity in *Salvia* (Lamiaceae): evidence from morphometrics and phylogenetic comparative methods. *Evolution* 74, 1335–1355. doi: 10.1111/evo.14030

Lara-Cabrera, S. I., Porter, J. M., and Steinmann, V. W. (in press). *The Freedom of Salvia s.l. and Resurrection of Lasemia and Ramona*. ALISO.

Larridon, I., Villaverde, T., Zuntini, A. R., Pokorny, L., Brewer, G. E., Epitawalage, N., et al. (2020). Tackling rapid radiations with targeted sequencing. *Front. Plant Sci.* 10:e01655. doi: 10.3389/fpls.2019.01655

Li, B., Cantino, P. D., Olmstead, R. G., Bramley, G. L. C., Xiang, C.-L., Ma, Z.-H., et al. (2016). A large-scale chloroplast phylogeny of the Lamiaceae sheds new light on its subfamilial classification. *Sci. Rep.* 6, 1–18. doi: 10.1038/srep34343

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Lopresti, A. L. (2017). *Salvia* (Sage): a review of its potential cognitive-enhancing and protective effects. *Drugs R D* 17, 53–64. doi: 10.1007/s40268-016-0157-5

Mandel, J. R., Barker, M. S., Bayer, R. J., Dikow, R. B., Gao, T.-G., Jones, K. E., et al. (2017). The compositae tree of life in the age of phylogenomics. *J. Syst. Evol.* 55, 405–410. doi: 10.1111/jse.12265

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10. doi: 10.14806/ej.17.1.200

Martínez-Gordillo, M. J., Bedolla-García, B., Cornejo-Tenorio, G., Fragoso-Martínez, I., García-Peña, M., del, R., et al. (2017). Lamiaceae de México. *Bot. Sci.* 95:780–806. doi: 10.17129/botsci.1871

Olvera-Mendoza, E. I., Bedolla-García, B. Y., and Lara-Cabrera, S. I. (2017). Revisión taxonómica de *Salvia* subgénero Calosphace sección Scorodoniae (Lamiaceae), endémica para México. *Acta Bot. Mex.* 118, 7–39. doi: 10.21829/abm118.2017.1198

Olvera-Mendoza, E. I., Godden, G. T., Montero-Castro, J. C., Porter, J. M., and Lara-Cabrera, S. I. (2020). Chloroplast and nuclear ribosomal cistron phylogenomics in a group of closely related sections in *Salvia* subg. *Calosphace*. *Braz. J. Bot.* 43:177–191. doi: 10.1007/s40415-019-00572-9

One Thousand Plant Transcriptomes Intitiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Qian, J., Song, J., Gao, H., Zhu, Y., Xu, J., Pang, X., et al. (2013). The complete chloroplast genome sequence of the medicinal plant *Salvia* miltiorrhiza. *PLoS ONE* 8:e57607. doi: 10.1371/journal.pone.0057607

R Core Team (2017). *R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing*. Vienna. Available online at: https://www.r-project.org/

Ramamoorthy, T. P. (1984). Notes on *Salvia* (Labiatae) in Mexico, with three new species. *J. Arnold. Arbor.* 65, 135–143.

Ramamoorthy, T. P., and Elliott, M. (1998). "Lamiaceae de México: diversidad, distribución, endemismo y evolución," in: *Diversidad biológica de México: orígenes y distribución*, eds T. P. Ramamoorthy, R. Bye, A. Lot, and J. Fa. (México; Instituto de Biología, UNAM), p. 501–525.

Rhodes, J. A., Baños, H., Mitchell, J. D., and Allman, E. S. (2021). MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R. *Bioinformatics*. 37:1766–1768. doi: 10.1101/2020.05.01.073361

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*. 74, 5463–5467. doi: 10.1073/pnas.74.12.5463

Santos, E. P. (1991). Genre *Salvia* L. Sous-genre calosphace (Benth.) Benth. Section nobiles (Benth.) Epl. (Labiatae). *Bradea* 4, 436–454.

Standley, P. C., and Williams, L. O. (1973). Flora of guatemala. *Fieldiana Bot.* 24, 273–301.

Straub, S. C. K., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., et al. (2011). Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. *BMC Genom.* 12:211. doi: 10.1186/1471-2164-12-211

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., and Liston, A. (2012). Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *Am. J. Bot.* 99, 349–364. doi: 10.3732/ajb.1100335

Torke, B. M. (2000). A revision of *Salvia* Sect. Ekmania (Lamiaceae). *Brittonia* 52, 265–302. doi: 10.2307/2666577

Turner, B. L. (2011). Recension of Mexican species of *Salvia* Sect. Standleyana (Lamiaceae). *Phytoneuron* 23, 1–6.

Villaverde, T., Jiménez-Mejías, P., Luceño, M., Waterway, M. J., Kim, S., Lee, B., et al. (2020). A new classification of Carex (Cyperaceae) subgenera supported by a HybSeq backbone phylogenetic tree. *Bot. J. Linn. Soc.* 194, 141–163. doi: 10.1093/botlinnean/boaa042

Villaverde, T., Pokorny, L., Olsson, S., Rincón-Barrado, M., Johnson, M. G., Gardner, E. M., et al. (2018). Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New Phytol.* 220, 636–650. doi: 10.1111/nph.15312

Walker, J. B. (2006). *A preliminary molecular phylogenetic analysis of Salvia subgenus Calosphace, Chap. 3* (PhD thesis). University of Wisconsin, Madison.

Walker, J. B., Drew, B. T., and Sytsma, K. J. (2015). Unravelling species relationships and diversification within the iconic california floristic province sages *Salvia* subgenus Audibertia, Lamiaceae). *Syst. Bot.* 40, 826–844. doi: 10.1600/036364415X689285

Walker, J. B., and Sytsma, K. J. (2007). Staminal evolution in the genus *Salvia* (Lamiaceae): molecular phylogenetic evidence for multiple origins of the staminal lever. *Ann. Bot.* 100, 375–391. doi: 10.1093/aob/mcl176

Walker, J. B., Sytsma, K. J., Treutlein, J., and Wink, M. (2004). *Salvia* (Lamiaceae) is not monophyletic: implications for the systematics, radiation, and ecological specializations of *Salvia* and tribe *Mentheae*. *Am. J. Bot.* 91, 1115–1125. doi: 10.3732/ajb.91.7.1115

Wanke, S., Granados Mendoza, C., Müller, S., Paizanni Guillén, A., Neinhuis, C., Lemmon, A. R., et al. (2017). Recalcitrant deep and shallow nodes in Aristolochia (Aristolochiaceae) illuminated using anchored hybrid enrichment. *Mol. Phylogenet. Evol.* 117, 111–123. doi: 10.1016/j.ympev.2017.05.014

Weitemier, K., Straub, S. C., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., et al. (2014). Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *App.l Plant Sci.* 2:1400042. doi: 10.3732/apps.1400042

Wells, T., Maurin, O., Dodsworth, S., Friis, I., Cowan, R., Epitawalage, N., et al. (2020). Combination of Sanger and target-enrichment markers supports revised generic delimitation in the problematic 'Urera clade' of the nettle family (Urticaceae). *Mol. Phylogenet. Evol.* 158:107008. doi: 10.1016/j.ympev.2020.107008

Wester, P., and Claßen-Bockhoff, R. (2011). Pollination syndromes of new world *Salvia* species with special reference to bird pollination. *Ann. Missouri Bot. Gard.* 98, 101–155. doi: 10.3417/2007035

Will, M., and Claßen-Bockhoff, R. (2017). Time to split *Salvia* s.l. (Lamiaceae) - new insights from old world *Salvia* phylogeny. *Mol. Phylogenet. Evol*. 109, 33–58. doi: 10.1016/j.ympev.2016.12.041

Wood, J. L. (2007). The *Salvias* (Lamiaceae) of Bolivia. *Kew Bull*. 62, 177–222.

Wu, Y. B., Ni, Z. Y., Shi, Q. W., Dong, M., Kiyota, H., and Gu, et al. (2012). Constituents from *Salvia* species and their biological activities. *Chem. Rev*. 14, 5967–6026. doi: 10.1021/cr200058f

Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., et al. (2012). FastUniq: a fast *de novo* duplicates removal tool for paired short reads. *PLoS ONE* 7:e52249. doi: 10.1371/journal.pone.0052249

Zeng, L., Zhang, N., Zhang, Q., Endress, P. K., Huang, J., and Ma, H. (2017). Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytol*. 214, 1338–1354. doi: 10.1111/nph.14503

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

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 November 2021
doi: 10.3389/fpls.2021.774232

Historical Dynamics of Semi-Humid Evergreen Forests in the Southeast Himalaya Biodiversity Hotspot: A Case Study of the *Quercus franchetii* Complex (Fagaceae)

*Si-Si Zheng[1,2], Xiao-Long Jiang[3], Qing-Jun Huang[2] and Min Deng[4,5]**

[1] Shanghai Chenshan Botanical Garden, Shanghai, China, [2] School of Ecological Technique and Engineering, Shanghai Institute of Technology, Shanghai, China, [3] The Laboratory of Forestry Genetics, Central South University of Forestry and Technology, Changsha, China, [4] School of Ecology and Environmental Science, Yunnan University, Kunming, China, [5] Yunnan Key Laboratory of Plant Reproductive Adaptation and Evolutionary Ecology, Yunnan University, Kunming, China

OPEN ACCESS

Edited by:
Susann Wicke, Humboldt University of Berlin, Germany

Reviewed by:
Christian Lampei, University of Münster, Germany
Tim Böhnert, University of Bonn, Germany

***Correspondence:**
Min Deng
dengmin.botany@gmail.com

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *11 September 2021*
Accepted: *28 October 2021*
Published: *23 November 2021*

Citation:
Zheng S-S, Jiang X-L, Huang Q-J and Deng M (2021) Historical Dynamics of Semi-Humid Evergreen Forests in the Southeast Himalaya Biodiversity Hotspot: A Case Study of the Quercus franchetii Complex (Fagaceae). Front. Plant Sci. 12:774232. doi: 10.3389/fpls.2021.774232

The Oligocene and Miocene are key periods in the formation of the modern topography and flora of East Asian and Indo-China. However, it is unclear how geological and climatic factors contributed to the high endemism and species richness of this region. The *Quercus franchetii* complex is widespread in the southeast Himalaya fringe and northern Indo-China with a long evolutionary history. It provides a unique proxy for studying the diversity pattern of evergreen woody lineages in this region since the Oligocene. In this study, we combined chloroplast (*cp*DNA) sequences, nuclear microsatellite loci (nSSRs), and species distribution modeling (SDM) to investigate the impacts of geological events on genetic diversity of the *Q. franchetii* complex. The results showed that the initial *cp*DNA haplotype divergence was estimated to occur during the middle Oligocene (30.7 Ma), which might have been raised by the tectonic activity at this episode to the Miocene. The nSSR results revealed two major groups of populations, the central Yunnan-Guizhou plateau (YGP) group and the peripheral distribution group when $K = 2$, in responding to the rapid YGP uplift during the late Miocene, which restricted gene flow between the populations in core and marginal areas. SDM analysis indicated that the distribution ranges of the *Q. franchetii* complex expanded northwards after the last glacial maximum, but the core distribution range in YGP was stable. Our results showed that the divergence of *Q. franchetii* complex is rooted in the mid-Oligocene. The early geological events during the Oligocene, and the late Miocene may play key roles to restrict seed-mediated gene flow among regions, but the pollen-mediated gene flow was less impacted. The uplifts of the YGP and the climate since LGM subsequently boosted the divergence of the populations in core and marginal areas.

Keywords: population genetic structure, ecological niche modeling, *Quercus* section *Ilex*, geoclimatic events, phylogeography

INTRODUCTION

The late Paleogene (36∼23.3 Ma) is a key period in the formation of the modern topography and flora of Asia (Akhmetiev and Zaporozhets, 2014; Li et al., 2019). During this period, an abrupt climate cooling at the Eocene-Oligocene (E-O) boundary (33.9 Ma) led to the turnover in regional biota and their distribution ranges. Meanwhile, the collisions between the Indian and Eurasian plates greatly changed the topography of Asia (Huchon et al., 1994; Chatterjee et al., 2013; Li S. H. et al., 2017). All these climatic and geological events had profound impacts on the distribution and divergence of the regional biota. However, little is known about how the timing and mechanisms of these ancient geological and climatic events that contributed to the high species diversity and high level of endemism in the southeast Himalayan fringe region.

One prevailing view addressed by numerous scholars is that the rapid uplifts of the Tibetan Plateau-Himalayas (TP) since the Miocene has created new habitats and niches, which have in turn promoted sympatric speciation (Liu et al., 2002; Wang et al., 2007; Meng et al., 2008). Meanwhile the uplift increased the complexity of the regional topography, which efficiently blocked gene flow among the populations, promoting allopatric speciation (Liu et al., 2006; Xu et al., 2010; Li et al., 2013; Qu et al., 2014). However, this view was challenged by Renner (2016) in a review that multiple line of evidences support TP had been 4–5 km high since the mid-Eocene, however, many phylogenetic works in Asia simply attributed the fast speciation between 0.5 and 15 Ma to the fast uplifts of TP, which is either miscited or outdated.

A recent study of Chinese angiosperms by Lu et al. (2018) indicated that the present Chinese flora is young, as a large number of genera did not originate until the Miocene (23 Ma), and their study determined that East China is a "museum" unlike West China, which is a "cradle" of herbaceous species. Likewise, the genome-wide analysis of *Salix brachista* (cushion willow) showed that the TP uplift induced sky island habitats, which increased population differentiation in combination with the Quaternary climate fluctuations, thus boosting *in situ* speciation (Chen et al., 2019). Other phylogenetic studies on relic woody species have also revealed recent divergence events, most dating back to the Miocene, with an intensification in the Pliocene-Pleistocene, regardless of whether these relic lineages had Paleogenic or even more ancient origins, e.g., *Cephalotaxus* (Cephalotaxaceae) (Wang et al., 2014), *Taxus* (Taxaceae) (Gao L. M. et al., 2007; Liu et al., 2013), and *Picea* (Pinaceae) (Shao et al., 2019).

Many study cases have collectively clarified the spatio-temporal diversity pattern of plant lineages in the East Himalayas since the Miocene. Indeed, the uplifts induced environmental heterogeneity and geographic barriers that together triggered the rapid diversification of these subtropical lineages. However, the ancient impacts of the geological events in the Oligocene to early Miocene remain a mystery, as the phylogeographic studies on the Paleogene diverged lineages are quite rare and fossil records in Asia at this epoch are scarce. Nevertheless, there are quite a few Paleogene relic genera also distributed in East Asian and the southwestern Himalayan fringe, e.g., *Ginkgo* (Shen et al., 2005; Gong et al., 2008), *Taxus* (Gao L. M. et al., 2007; Ji et al., 2020), and *Eurycorymbus* (Wang et al., 2009). Phylogeographic studies on those relic species have illustrated that southwestern China, Dabashan, and the Wuyi Mountain regions served as important refugia during time periods with extreme climates (Shen et al., 2005; Gao L. M. et al., 2007; Wang et al., 2009). However, most of these lineages suffered massive extinctions at the geological timescale, with only few extant species with rather restricted distributions remaining. Therefore, investigations of their spatial genetic pattern can only provide limited information about the evolutionary dynamics occurring further back in geological time.

Quercus franchetii and *Q. lanata* belong to *Quercus* section *Ilex*. The two species are widespread in southwestern China and the southern Himalayas to Northern Indo-China, respectively, at an elevation range of approximately 800–2,600 m (Govaerts and Frodin, 1998; Huang et al., 1999). Their distribution ranges cover the main area of the ancient Red River drainage basin. They are both key regional trees in semi-humid evergreen broad-leaved forests and dry-hot river valleys and have important ecological service and functions (Wangda and Ohsawa, 2006; Liu et al., 2011, 2012). The early derived status of *Q. franchetii* in section *Ilex* at the E-O boundary was inferred by recent phylogenetic studies on oaks (Jiang et al., 2019; Hipp et al., 2020). Our recent phylogenetic study on *Quercus* section *Ilex* supplemented *Q. lanata* for analysis. The result showed that *Q. lanata* and *Q. franchetii* are sister taxa (unpublished data). However, after adding RAD-seq data from more individuals from both species and reconstructing the phylogenetic tree, neither of the two species form a monophyletic clade (unpublished data). Likewise, morphometric measurements showed no difference between *Q. franchetii* and *Q. lanata*, both at the species level as well as among different geographical regions (Zheng, 2021). These results together suggested that the two species are very closely related to each other or indeed represent the same species (hereafter called "*Quercus franchetii* complex"). Data beyond the molecular dating results concur that the *Q. franchetii* complex has a long evolutionary history. Fossils resembling the extant *Q. franchetii* complex have been widely reported along the Tethys/Paratethys Seaway dating to the late Eocene to Pliocene, and they were commonly used as a proxy indicating warm and semi-humid climates (Bouchal et al., 2017; Denk et al., 2017; Guner et al., 2017). Thus, the *Q. franchetii* complex offers a unique opportunity for untangling the timing and possible mechanisms by which geological events since the Oligocene have shaped the high biodiversity and endemism level of the southeastern Himalaya fringe.

In this study, we comprehensively sampled populations of the *Q. franchetii* complex throughout its distribution range (**Figure 1**). We used *cp*DNA and nSSR markers to scan the populations. By coupling population genetic structure analyses with species distribution modeling (SDM), we aimed to (1) illustrate the spatial genetic structure of the *Q. franchetii* complex, (2) identify the key environmental factors restricting the distribution and diversity pattern of the species complex, and (3) explore the key factors that drove the divergence of the *Q. franchetii* complex. This study provides deep insights into the distribution and evolutionary dynamics of this subtropical

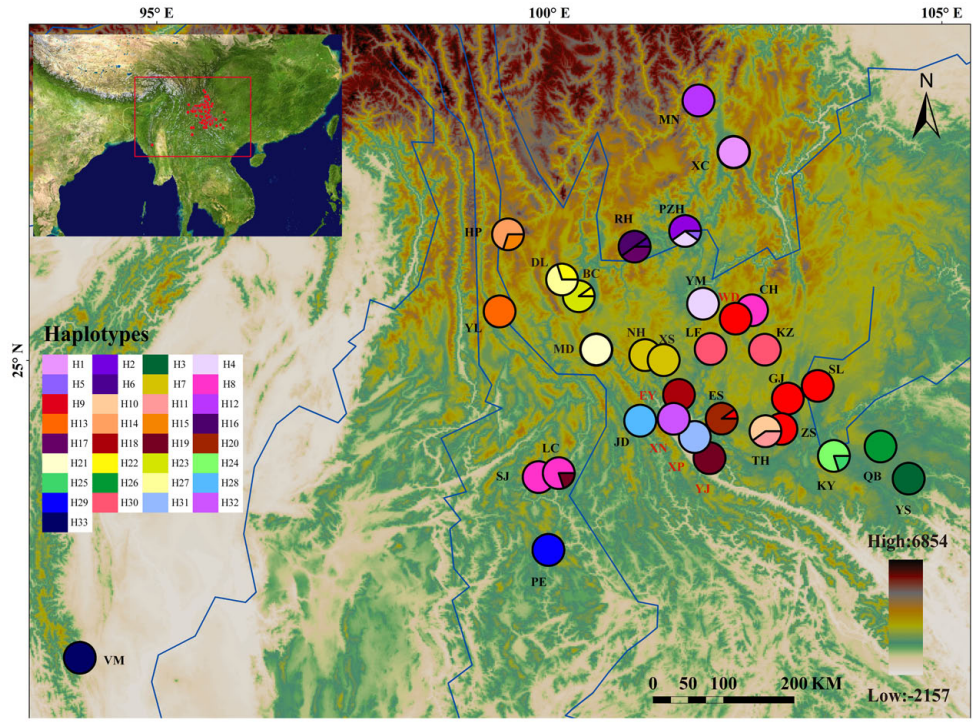


FIGURE 1 | Geographic distribution of 33 *cp*DNA haplotypes detected in the *Quercus franchetii* complex. The colored pie charts representing the frequencies of haplotypes at each sampling site. Haplotype colors corresponding to the charts are shown in the left panel. The color scale representing different elevation gradients are shown in the lower-right panel. The population with less than 10 individuals were marked with red population labels.

woody lineage in the context of global environment change, informing efforts to safeguard this unique forestry ecosystem in the Southeast Himalaya biodiversity hotspot.

MATERIALS AND METHODS

Ethics Statement

Sampling of oaks were granted and supported by National Forestry Bureau of China, Local National Nature Reserves, and Ministry of Environmental Conservation and Forestry, Myanmar.

Population Sampling

Thirty-three *Q. franchetii* complex populations were sampled from Yunnan and Sichuan, China and Mount Victoria in Chin State, Myanmar. In total, 303 individuals from 33 populations were sampled for this study, covering the major known distribution range of the *Q. franchetii* complex (**Table 1**). Samples from the same population came from individuals that were separated by at least 50 m. At least 10 trees were sampled from each population, except for populations with very few individuals (WD, EY, YJ, XP, XN), in which case we sampled all the accessible adult trees in those populations. Fresh and healthy mature leaves were collected and put into containers with silica gel to dry them quickly until DNA extractions could be performed. The voucher specimens of the DNA samples were deposited in the herbarium of Shanghai Chenshan Botanical Garden (CSH).

DNA Extraction, PCR Amplification, and Sequencing

Total genomic DNA was extracted using a modified cetyltrimethyl ammonium bromide (CTAB) protocol (Doyle and Doyle, 1987). Three pairs of *cp*DNA primers, namely *psb*A-*trn*H (Shaw et al., 2005), *trn*T-*trn*L (Taberlet et al., 1991), and *atp*I-*atp*H (Grivet et al., 2001), and eight highly polymorphic nuclear microsatellite loci, specifically nuclear simple sequence repeats (nSSRs), were selected for genotyping on all samples. Primer sequences and PCR amplification conditions are summarized in **Supplementary Appendix 1**. The *cp*DNA and nSSR amplification conditions followed the methods described by Xu et al. (2015) and An et al. (2016), respectively. PCR products of nSSR markers were 10 times diluted, then mixed with fluorescence size standards at a ratio of 6-FAM: HEX: ROX = 1:1:2, then genotyped by Shanghai Majorbio Bio-pharm Technology Co., Ltd. (Shanghai, China). The PCR products of *cp*DNAs were purified, then biodirectionally sequenced by the same company. All sequences obtained in this study have been deposited in GenBank (see "Data Availability").

The *cp*DNA sequences were assembled and checked using Sequencher 4.01 (Gene Codes Corp., Ann Arbor, MI, United States). The ClustalW implementation in MEGA X (Kumar et al., 2018) was used for sequence alignment with

TABLE 1 | Sampling information, nSSR and *cp*DNA genetic diversity, probabilities of populations belonging to each genetic cluster (C_A, C_B, C_C, C_D) inferred by InStruct analyses, locality habitat suitability, stability obtained from SDMs, and bottleneck effect test for the *Quercus franchetii* complex.

Species	Pop code	Lon	Lat	n	*cp*DNA			SSRs						SDM			Bottleneck
					Haplotypes (no. of individuals)	*h*	$\pi \times 10^3$	C_A	C_B	C_C	C_D	*Ar*	*He*	N_{Pre}	N_{LGM}	N_{stab}	P_W_2t
QF	XC	102.35	27.66	10	H1 (10)	0	0	0.008	0.900	0.083	0.009	3.79	0.532	0.215	0.224	0.991	**0.020***
QF	PZH	101.73	26.65	10	H2 (6), H4 (3), H5 (1)	0.6	0.042	0.042	0.928	0.018	0.012	3.83	0.529	0.703	0.457	0.754	**0.012***
QF	LF	102.05	25.15	10	H30 (10)	0	0	0.038	0.028	0.932	0.006	3.85	0.557	0.778	0.686	0.909	0.313
QF	YM	101.96	25.72	10	H4 (10)	0	0	0.025	0.937	0.023	0.015	4.17	0.612	0.786	0.655	0.869	0.547
QF	RH	101.08	26.45	10	H6 (1), H16 (5), H17 (4)	0.64	0.028	0.011	0.964	0.009	0.016	3.55	0.519	0.757	0.471	0.714	0.250
QF	NH	101.22	25.07	10	H7 (10)	0	0	0.031	0.048	0.914	0.008	3.96	0.569	0.754	0.627	0.873	1.000
QF	ZXS	101.46	25.01	10	H7 (10)	0	0	0.87	0.033	0.08	0.017	3.87	0.564	0.731	0.598	0.867	0.688
QF	CH	102.58	25.64	10	H8 (10)	0	0	0.141	0.066	0.787	0.006	3.85	0.573	0.688	0.579	0.891	1.000
QF	SJ	99.86	23.51	10	H8 (10)	0	0	0.957	0.017	0.02	0.006	3.23	0.486	0.253	0.391	0.862	0.641
QF	LC	100.12	23.57	10	H8 (8), H19 (2)	0.36	0.011	0.687	0.022	0.286	0.005	3.18	0.518	0.391	0.323	0.932	0.938
QF	GJ	103.04	24.52	10	H9 (10)	0	0	0.978	0.006	0.01	0.006	3.47	0.523	0.760	0.637	0.878	0.383
QF	ZS	102.95	24.13	10	H9 (10)	0	0	0.954	0.008	0.032	0.006	3.38	0.486	0.767	0.658	0.891	0.313
QF	SL	103.42	24.68	10	H9 (10)	0	0	0.073	0.876	0.042	0.009	3.66	0.563	0.627	0.572	0.945	0.313
QF	WD	102.37	25.53	1	H9 (1)	1	0	0.78	0.006	0.209	0.005	–	–	0.684	0.550	0.867	–
QF	ES	102.19	24.26	10	H9 (1), H20 (9)	0.2	0.013	0.932	0.028	0.018	0.023	4.02	0.586	0.707	0.723	0.983	0.055
QF	TH	102.75	24.10	10	H10 (6), H11 (4)	0.53	0.028	0.904	0.022	0.061	0.014	3.46	0.523	0.736	0.635	0.899	0.313
QF	MN	101.90	28.31	10	H12 (10)	0	0	0.009	0.946	0.02	0.025	3.85	0.557	0.509	0.234	0.725	0.945
QF	YL	99.37	25.63	10	H13 (10)	0	0	0.028	0.928	0.04	0.009	3.5	0.536	0.520	0.305	0.785	0.461
QF	HP	99.47	26.61	10	H14 (7), H15 (3)	0.47	0.005	0.007	0.971	0.014	0.009	3.79	0.544	0.386	0.164	0.778	0.250
QF	EY	101.65	24.56	6	H18 (6)	0	0	0.781	0.184	0.028	0.006	3.5	0.568	0.770	0.687	0.917	0.945
QF	YJ	102.04	23.76	7	H19 (7)	0	0	0.841	0.029	0.119	0.011	3.23	0.437	0.638	0.631	0.993	**0.020***
QF	MD	100.60	25.14	10	H21 (10)	0	0	0.064	0.024	0.898	0.015	3.39	0.506	0.764	0.643	0.879	0.375
QF	BC	100.37	25.82	10	H22 (1), H23 (9)	0.2	0.026	0.018	0.019	0.958	0.005	3.69	0.530	0.703	0.486	0.783	0.195
QF	DL	100.16	26.03	10	H22 (3), H27 (7)	0.47	0.005	0.808	0.031	0.122	0.039	4.34	0.604	0.653	0.388	0.735	0.074
QF	KY	103.62	23.78	10	H24 (8), H25 (2)	0.36	0.004	0.029	0.772	0.174	0.025	4.01	0.588	0.766	0.667	0.902	0.461
QF	KZ	102.74	25.14	10	H30 (10)	0	0	0.108	0.102	0.765	0.025	3.67	0.540	0.681	0.559	0.879	0.313
QF	XN	101.58	24.26	3	H32 (3)	0	0	0.047	0.021	0.927	0.005	–	–	0.350	0.759	0.591	–
QL	YS	104.58	23.49	10	H3 (10)	0	0	0.925	0.046	0.011	0.018	3.73	0.579	0.465	0.574	0.891	1.000
QL	JD	101.16	24.23	12	H28 (12)	0	0	0.020	0.018	0.942	0.02	3.46	0.549	0.651	0.726	0.924	0.641
QL	QB	104.22	23.91	10	H26 (10)	0	0	0.022	0.94	0.031	0.007	3.95	0.603	0.600	0.565	0.965	0.742
QL	XP	101.85	24.03	3	H31 (3)	0	0	0.868	0.057	0.064	0.011	–	–	0.699	0.713	0.986	–
QL	PE	99.99	22.59	10	H29 (10)	0	0	0.007	0.026	0.014	0.953	3.28	0.522	0.151	0.132	0.981	0.297
QL	VM	94.02	21.21	11	H33 (11)	0	0	0.009	0.009	0.01	0.972	3.7	0.586	0.118	0.074	0.955	0.844

QF, Q. franchetii; QL, Q. lanata; Lon, longitude; Lat, latitude; n, number of individuals investigated in the population; h, haplotype diversity; π, nucleotide diversity; H_e, expected heterozygosity; A_r, standardized allelic richness; N_{Pre}, present habitat suitability; N_{LGM}, last glacial maximum (LGM) habitat suitability; N_{stab}, habitat stability since the LGM; P_W_2t, P-values of the TPM model based on the Wilcoxon sign-rank test; (the bold values), significant correlation ($P < 0.05$).*

manual adjustment. All microsatellite loci were checked using GeneMarker® (Hulce et al., 2011) to detect and analyze allele sizes. Null alleles and stutter bands were checked with MicroChecker (Van Oosterhout et al., 2004).

Genetic Diversity and Structure

Haplotypes and polymorphism statistics for *cp*DNA loci were calculated with DnaSP 6.0 (Rozas et al., 2017). The haplotype geographic distribution was projected onto a map using ArcGIS 10.5.[1] Total haplotype diversity (H_T), within-population diversity (Hs), and coefficients of differentiation (G_{ST} and N_{ST}) for *cp*DNA loci were estimated using PERMUT 2.0 (Pons and Petit, 1996). Haplotype diversity (h) and nucleotide diversity (π) for *cp*DNA loci were obtained with ARLEQUIN 3.5 (Excoffier and Lischer, 2010). The *cp*DNA haplotype network was constructed using the Median-Joining model in NETWORK 5.0.0.1 (Bandelt et al., 1999). GenAlEx 6.5 (Peakall and Smouse, 2012) was used to calculate the genetic diversity index of microsatellite data, including expected heterozygosity (H_e), observed heterozygosity (H_o), number of alleles (N_A), and effective number of alleles (N_E). Allelic richness (A_r) of nSSRs was determined using HP-RARE (Kalinowski, 2005).

The genetic differentiation values (F_{ST}) based on *cp*DNA and nSSR data were estimated using ARLEQUIN 3.5, respectively

[1] http://www.esri.com/software/arcgis/

(Excoffier and Lischer, 2010). The Genetic Landscape GIS toolbox (Vandergast et al., 2011) in ArcGIS 10.5 was used to generate a geographical landscape map based on both genetic diversity (A_r) for *cp*DNA loci and genetic divergence (F_{ST}) based on *cp*DNA and nSSR data according to inverse distance weighted interpolation. For *cp*DNA and nSSR data, analysis of molecular variance (AMOVA) was performed to assess the genetic variation among populations and within populations. The Mantel test was also performed based on the genetic structure and geographic distance matrix with 1,000 random permutations to evaluate their relationship and test isolation by distance (IBD). For nSSR data, Principal coordinate analysis (PcoA) based on genetic distance was performed using GenAlex 6.5 (Peakall and Smouse, 2012) to assess differences among individuals or groups. Barrier (Manni et al., 2004) was used to set up geographic barriers according to sample locations to detect the existence of genetic barriers among populations.

We used GenePop v 4.2 (Rousset, 2008) to test departures from Hardy-Weinberg equilibrium (HWE) for each of the eight nSSR loci. As the nSSR loci significantly deviated from HWE (see Results section "Genetic Diversity and Structure"), Bayesian assignment probability methods using the programs InStruct (Gao H. et al., 2007) and STRUCTURE 2.3.4 (Pritchard et al., 2000) were both used to infer the population genetic structure. The number of clusters (K) was varied from 1 to 10, with 10 replicates at each value. Each run consisted of a burn-in length of 25,000 iterations with a run length of 500,000 MCMC (Markov chain Monte Carlo) iterations. The optimal K (number of clusters) was determined using the $\triangle K$ method (Evanno et al., 2005). CLUMPP (Jakobsson and Rosenberg, 2007) was used to align 10 runs of InStruct with the optimum K using a greedy algorithm.

Among the three models in the Bottleneck program (Luikart et al., 1998), the "Two-phase mutation model" (TPM) is the most suitable for microsatellite loci, which was thus selected to test whether the populations of the *Q. franchetii* complex had experienced a bottleneck. We used "Wilcoxon sign-rank test" method for a significance test.

Divergence Time Estimation

The divergence time dating for *cp*DNA haplotypes of the *Q. franchetii* complex was estimated using a Bayesian approach as implemented in BEAST V2.4 (Suchard et al., 2018). *Quercus glauca* (*Quercus* section *Cyclobalanopsis*) was chosen as the outgroup to root the tree. An uncorrelated lognormal relaxed clock was applied with the K81uf + I substitution model, which was selected based on the Akaike information criterion (AIC) in Modeltest 3.7 (Posada and Crandall, 1998). The earliest conclusive leaf fossils of *Quercus* section *Ilex* were discovered in Tibet, dated to the late Eocene (ca. 34 Ma) (Su et al., 2019) was set as the minimum age to constrain the stem of the haplotype tree of the *Q. franchetii* complex with a lognormal distribution and a median of 37.8 Ma (95% HPD: 34.02–56.6 Ma). The MCMC chains were run for 100 million generations with a sampling frequency of once every 10,000 generations. Convergence was assessed using Tracer v1.7[2] (Rambaut et al., 2018), and the effective sample sizes for all parameters were calculated. The resulting tree and log files from the two replicate runs were combined with LogCombiner v1.8. Then, we used TreeAnnotator v. 1.8[3] to generate the maximum clade credibility (MCC) tree after discarding the first 20% of the trees as burn-in. The results were visualized using FigTree v1.4.3.[4]

Population Demographic History and Ancestral Area Reconstruction Analyses

Pairwise mismatch distribution analysis for *cp*DNA loci, Tajima's (1989) D and Fu's (1997) Fs of neutrality tests were performed to detect possible demographic expansions of the *Quercus franchetii* complex using ARLEQUIN 3.5 (Excoffier and Lischer, 2010).

Ancestral range reconstruction was analyzed using statistical dispersal-vicariance (S-DIVA) analysis as implemented in RASP (Yu et al., 2015) and DEC models (Ree et al., 2005; Ree and Smith, 2008). Four areas were delimited for ancestral area reconstruction based on the geological characteristics and biogeographical division of southwestern China, including (A) the Nanpan River region (NPR); (B) the southwestern Red River (including southern Himalayas to Red River) (RR); (C) the Hengduan Mountains (HDM), and (D) the Yunnan-Guizhou Plateau (YGP) (**Figure 2A**).

Species Distribution Modeling

We used MaxEnt 3.4 (Phillips et al., 2006) to simulate the potential distribution range of the *Q. franchetii* complex (including *Q. franchetii* and *Q. lanata*) under the past, current, and projected future climate scenarios based on the maximum entropy model (Phillips et al., 2004). In total, 145 accurate occurrence points were collected from the Global Biodiversity Information Facility (GBIF),[5] Chinese Digital Herbarium (CVH),[6] and the field collection records of our research team. Each voucher specimen in the distribution record was inspected and checked carefully. Nineteen bioclimatic variables with a 2.5-arc-min resolution for the present (1950–2000), the last glacial maximum (LGM) (CCSM) period, and the future (2060–2080, RCP2.6; RCP4.5; RCP6.0; RCP8.5) were downloaded from WorldClim 2.[7] The nine environmental factors (bio1, Annual Mean Temperature; bio4, Temperature Seasonality; bio6, Min Temperature of Coldest Month; bio7, Temperature Annual Range; bio10, Mean Temperature of Warmest Quarter; bio11, Mean Temperature of Coldest Quarter; bio12, Annual Precipitation; bio13, Precipitation of Wettest Month; bio17, Precipitation of Driest Quarter) were selected after eliminating climatic variables such that none that were include were highly correlated (i.e., with correlation coefficients

[2]http://tree.bio.ed.ac.uk/software/tracer/
[3]http://beast.bio.ed.ac.uk/TreeAnnotator
[4]http://tree.bio.ed.ac.uk/software/figtree/
[5]http://www.gbif.org
[6]https://www.cvh.ac.cn
[7]http://www.worldclim.org/

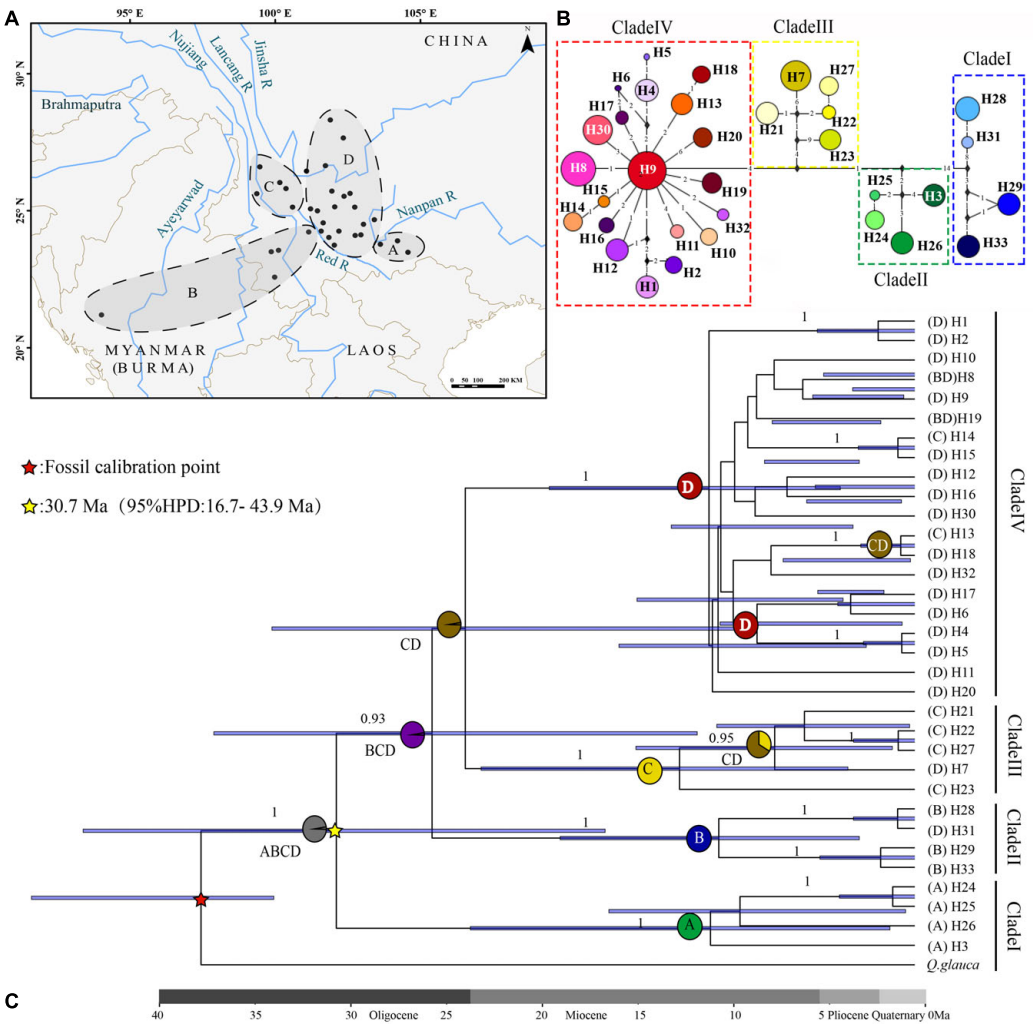


FIGURE 2 | Phylogenetic analysis and ancestral distribution area reconstruction of the *Quercus franchetii* complex. **(A)** Ancestral distribution area of the *Q. franchetii* complex. A, Nanpan River region (NPR); B, southwestern Red River (RR); C, Hengduan Mountains (HDM); D, Yunnan–Guizhou Plateau (YGP). **(B)** Haplotypes of the *cp*DNA network of the *Quercus franchetii* complex. The colors in the pie charts represent different *cp*DNA haplotypes. Numbers on the branches indicate the number of substitutions. Black diamonds indicate unsampled or extinct ancestral haplotypes. **(C)** BEAST-derived chronogram of the *Q. franchetii* complex. The blue bar length represents the 95% HPD of the species divergence time, and circles at nodes represent the distribution range. Values below branches represent the posterior probability.

greater than 0.8) using the Dismo package in R.[8] We also chose CHELSA climate data,[9] which had a higher prediction power in mountain regions (specifically the Himalayas), to simulate the present distribution (Bobrowski and Schickhoff, 2017; Karger et al., 2017).

In order to improve the prediction accuracy of the model, it was necessary to optimize the model in MaxEnt 3.4 (Phillips et al., 2006) by setting the β multiplier and environmental characteristic parameters. The MaxEnt model captures five features: linear (L), quadratic (Q), hinge (H), product (P), and threshold (T). In this research, we used seven feature combinations (auto, L, H, LQ, LPQ, LQH, LQHP) and set the regularization multiplier from 0.5 to 10 with increments of 0.5. Then, we used ENMTools (Warren et al., 2010) to calculate the lambdas file for the maxent result, and selected the model with the smallest AIC value as the optimal model parameter for further analyses. The occurrence points of *Q. franchetii* and *Q. lanata* were randomly divided into 75 and 25% of the data for training and testing, respectively. The area under the curve (AUC) was used to evaluate the model's accuracy. AUC values ranged from 0.5 to 1, where the higher AUC value indicated a better prediction. The maximum training sensitivity plus specificity thresholds for the presence or absence of species was used to draw a species distribution map in ArcGIS 10.5. In addition, we classified three categories equally between this threshold and 1, corresponding to low, medium, and high fitness areas, respectively. To compare the changes in species distribution across different periods, three indicators were

[8]https://CRAN.R-project.org/package=dismo

[9]https://chelsa-climate.org/

calculated: locality habitat stability (N_{Stab}), habitat distribution area ratio (N_a), and habitat expansion extent (N_e). These values were calculated using the following formulas: $N_{Stab} = 1- \mid N_{Pre} - N_{LGM} \mid$, where N_{Pre} and N_{LGM} are the habitat suitability of the present and LGM distribution area; N_a = (present distribution areas)/(LGM or future distribution areas): a value close to 1 indicates a stable distribution of the species, while a value much higher or lower than 1 indicates that the distribution area of the species has expanded or contracted from the LGM to the present; N_e = [1 - (Distribution area overlapping between the LGM and present or the future and present/present distributions area)] × 100% represents the percentage of the distribution that has expanded from the LGM to the present.

The potential dispersal routes of the *Q. franchetii* complex in the past and present periods were inferred based on the least-cost path analysis method using SDM toolbox 2.0 (Brown, 2014) in ArcGIS 10.5. The specific steps are listed below: Firstly, we generated a resistance layer by inverting the SDMs (1-SDM). The resistance layer was used to create a cost distance raster for each sample locality. The corridor layers were built between two locations that only share haplotypes based on the cost distance raster. We used the categorical the least cost path (LCP) approach to better describe the habitat heterogeneity and its role in the dispersal. The value of each corridor layer was divided into low, medium, and high, and then these three intervals were re-divided into new values 5, 2, and 1. Finally, we reclassified all the corridor layers, summarized and standardized them from 0 to 1, and determined the dispersal corridors of the *Q. franchetii* complex.

Detection of Correlations Between Genetic Diversity and Climatic Factors

The linear model in R 3.5[10] was used to estimate the correlation of genetic diversity indexes (A_r and H_e) and genetic structure (cluster A from the Bayesian clustering, C_A) with habitat and geographic factors. Five variables—population longitude and latitude, habitat suitability for the present (N_{Pre}) and the LGM (N_{LGM}), and habitat stability (N_{Stab}) were used as explanatory covariates.

RESULTS

Genetic Diversity and Structure

The *psb*A-*trn*H, *atp*I-*atp*H, and *trn*T-*trn*L sequence alignments were 482–520, 899–1030, and 796–838 bp in length, respectively. The combined length of the three aligned chloroplast fragments was 2,319 bp, with 94 polymorphisms across a total of 33 haplotypes among the 303 individuals analyzed. The *cp*DNA haplotype diversity was h = 0.956 and nucleotide diversity was π = 0.00536. The nucleotide diversity and haplotype diversity within the populations were 0–0.04184 and 0–0.644, respectively (**Table 1** and **Figure 1**). The total diversity (H_T = 0.982) was much higher than the average within-population diversity (H_S = 0.123). The genetic diversity map showed that the NPR and HDM populations had high genetic diversity (**Figure 3C**). N_{ST} (0.959) was significantly greater than G_{ST} (0.874) ($P < 0.05$), which indicated that a clear phylogeographic structure existed among the *Q. franchetii* complex populations. The network analysis resolved four distinct *cp*DNA haplotype clades in the *Q. franchetii* complex, which was consistent with the haplotype structure determined by BEAST. The haplotypes in Clade IV exhibited a star-like structure. The most common haplotype, H9, had the widest distribution, mainly within the central YGP (**Figure 2B**).

Among the 303 individuals genotyped, a total of 115 alleles were identified using eight pairs of microsatellite primers. All eight microsatellite loci were highly polymorphic, with allele numbers varying from 8 to 18 per locus. The genetic diversity indexes N_A, N_E, H_e, and H_o were 4.182 (SE = 0.103), 2.578 (SE = 0.066), 0.537 (SE = 0.013), and 0.536 (SE = 0.016), respectively, while allele richness (A_r) was 3.18–4.34.

The AMOVA on *cp*DNA sequence data revealed greater genetic variation among populations (87%; F_{ST} = 0.87) than within populations (13%). In contrast, the nSSRs showed substantial genetic differences within populations (63%; F_{ST} = 0.369, **Table 2**). Pairwise F_{ST} values calculated based on chloroplast and SSR data were shown in **Supplementary Appendix 2**. In order to more visually observe genetic differentiation, F_{ST} values were projected onto a map and a genetic divergence map was made. The genetic divergence map showed that the population in southwestern Yunnan had a relatively high genetic divergence, while the divergence of the populations from the northeast was comparatively lower (**Figures 3A,B**). The Mantel tests of *cp*DNA data (r = 0.052, P = 0.001) and microsatellite data (r = 0.413, P = 0.001) both revealed significant correlations between genetic and geographic distances, but the correlations among chloroplast markers were weaker (**Supplementary Appendix 3**). Barriers inferred from nuclear genes showed a barrier between the YGP and HDM regions (**Figure 4**). As well as, distinct geographical isolation had been detected between the populations in southwestern Yunnan and NPR (**Figure 4**). Based on the TPM model, three populations were determined to have experienced bottleneck (PZH, XC, YJ; **Table 1**).

A total of three loci of nSSRs conformed to HWE, and five loci deviated from HWE ($P < 0.05$, **Supplementary Appendix 1**). In the Bayesian clustering analysis, the optimal K-value for STRUCTURE was K = 4 (**Supplementary Appendix 4**). However, the optimal K-value selected by InStruct was 2, and the second highest peak occurred for K = 4 (**Supplementary Appendix 4**). Thus, we performed cluster analysis on the *Q. franchetii* complex using K = 2 and K = 4, respectively for both STRUCTURE and InStruct. Following the InStruct runs with K = 2, cluster a was mainly composed of the YGP group, with some populations in the HDM (DL, BC, MD), and southwestern Yunnan (LC, SJ) groups. Cluster b was composed of 12 populations peripheral to those of cluster a. When K = 4, the YGP group was further divided into subgroups, and the populations from RR (PE, VM) separated from the rest of cluster b to form a new subgroup (**Figure 4**). The STRUCTURE results were identical to the results obtained from STRUCTURE (**Supplementary Appendix 5**). The principal coordinate analysis (PcoA) of the clustering results found that 33 natural populations

[10]http://www.r-project.org/

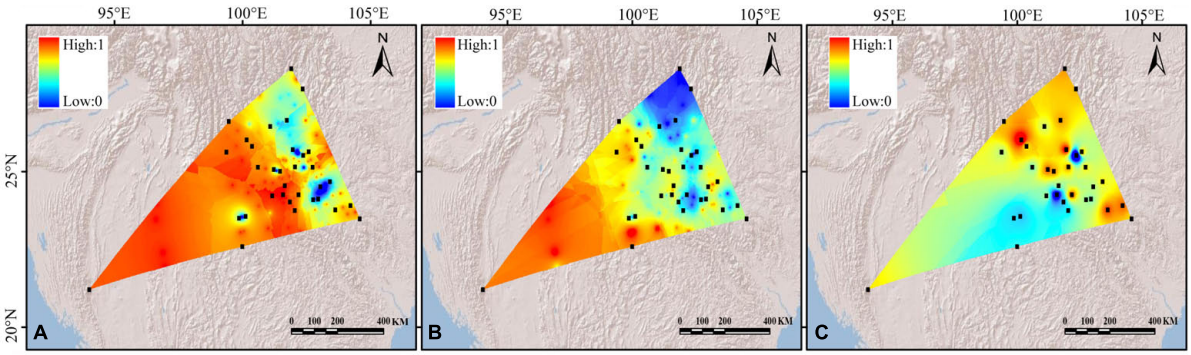


FIGURE 3 | Spatial interpolation of genetic differentiation (*F*st) **(A)** *cp*DNAs, **(B)** nSSRs, and **(C)** genetic diversity of the *Quercus franchetii* complex based on *cp*DNAs. The color ranges from blue to red, representing the genetic differentiation or genetic diversity values from low to high.

TABLE 2 | Molecular variance analysis of the *Quercus franchetii* complex.

	Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation%
***cp*DNA**	Among populations	31	126.684	0.427	87%
	Within populations	270	17.200	0.064	13%
	Total	301	143.884	0.491	100%
SSR	Among populations	31	1014.597	2.939	37%
	Within populations	270	1356.910	5.026	63%
	Total	301	2371.507	7.965	100%

can be divided into four regions, which is mostly consistent with the results obtained using InStruct. The first and second principal components explained 16.24 and 6.91% of the genetic variation, respectively (**Supplementary Appendix 6**).

Divergence Time Estimation

The crown age of *cp*DNA haplotypes in the *Q. franchetii* complex was dated to the late Oligocene (30.7 Ma, 95% highest posterior density, HPD = 16.7–43.9 Ma), and Clade I (NPR) was the earliest derived. The haplotypes of the southwestern lineage (Clade II) began to diversify around 25.7 Ma (95% HPD = 11.9–37.2 Ma). The divergence of Clades III (HDM region) and IV (YGP region) was dated to ca. 24 Ma (95% HPD = 9.95–34.13 Ma), with subsequent rapid divergence of the haplotypes during the late Miocene (**Figure 2**).

Demographic History and Ancestral Area Reconstruction

Demographic History

Neutrality tests (Tajima's $D = -0.50558$, $P = 0.325$; Fu's $Fs = 2.77576$, $P = 0.789$) failed to identify population expansion. The mismatch distribution for *cp*DNA showed a multimodal distribution, consistent with a stable population size (**Supplementary Appendix 7**).

Ancestral Area Reconstruction

The ancestral distribution range reconstruction of the S-DIVA and DEC models both showed that the *Q. franchetii* complex had a wide distribution of ancestral populations. The S-DIVA model (95% HPD = 99%) provided a higher confidence estimate than did the DEC model (95% HPD = 45.25%), and its results indicated that the ancestral distribution area of the *Q. franchetii* complex was widespread in southwestern Yunnan and the Southern Himalayas (including A, B, C, and D). Then followed three vicariance events, which led to the divergences of the NPR lineage (Clade I) during the Oligocene (95% HPD = 99%) and the RR lineage (Clade II) during the late Oligocene episode (95% HPD = 98.97%), with an increase in the divergence of the HDM lineage (Clade III) and YGP lineage (Clade IV) during the early Middle Miocene (**Figure 2**). The DEC results are shown in **Supplementary Appendix 8**.

Ecological Niche Modeling

Ecological niche modeling of the *Q. franchetii* complex in Maxent using WorldClim climate data revealed a high performance score (AUC = 0.9679–0.9719, standard deviation = 0.0157). Annual Mean Temperature (bio1) was the greatest contributor (36.46%, standard deviation = 1.3), followed by Temperature Seasonality (bio4) (34.7%, standard deviation = 1.3) and Annual Precipitation (bio12) (7.72%, standard deviation = 1.52) in identifying the areas of occurrence for *Q. franchetii* complex populations. The maximum sensitivity plus specificity value 0.13 was used as the species absence/presence threshold. The current distribution of the *Q. franchetii* complex was similar to the predicted distribution, except for quite a few occurrence sites in southwestern and southeastern Yunnan and Southern Himalayas that were located in predicted unsuitable areas (**Figure 5B**). The potential distribution range retreated to the south during the LGM (**Figure 5A**). The total distribution area of the LGM was greater than that of the present (N_a = 0.42,

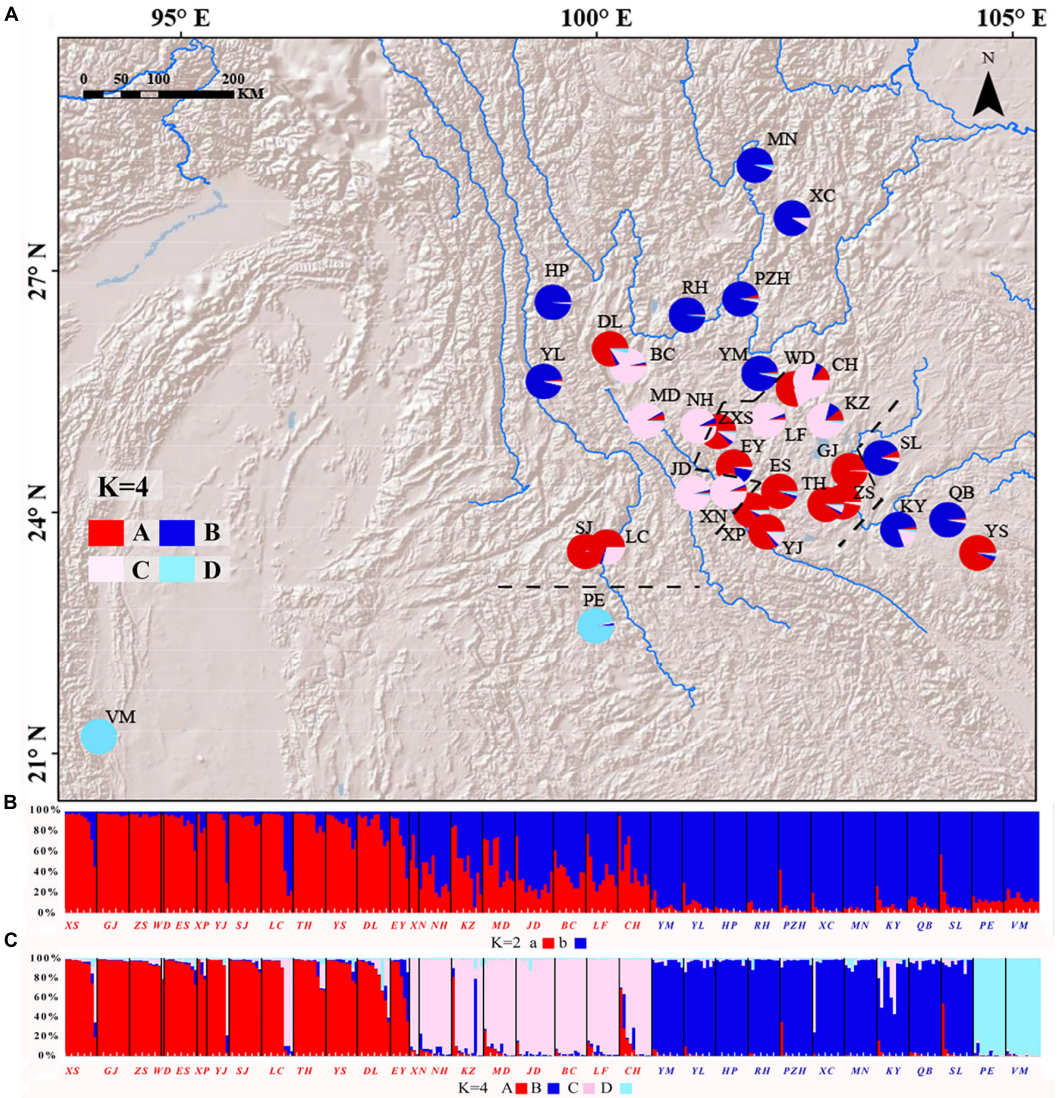


FIGURE 4 | (A) Geographic distribution of the *Quercus franchetii* complex according to STRUCTURE grouping analyses. STRUCTURE cluster analysis diagram when **(B)** K = 2 and **(C)** K = 4. The colors in the pie charts represent different groupings, and the black dotted line indicates the inferred geographic isolation based on Barrier.

N_e = 21%). In the future period, the higher the RCP value was, the more obvious of an expansion in the suitable area to the northwest was detected. The ratios of the present range to the future range under RCP2.6, RCP4.5, RCP6.0, and RCP8.5 were 1.07, 1.15, 1.04, and 1.18, respectively. Accordingly, the highly suitable habitat would be reduced by 7.2, 7.2, 16.5, and 40.1%, respectively (**Figures 5C–F**). The predicted present distribution of *Q. franchetii* complex using CHELSA climate data was similar to that from WorldClim climate data (**Supplementary Appendix 9**). The suitable distribution area in the current period from the two data sources was largely overlapped (0.94). In order to better compare the distribution dynamics with other oak, of which WorldClim database were generally used for distribution simulation, we selected the results of WorldClim for the subsequent analyses.

Putative dispersal corridors in the two periods were visualized based on *cp*DNA haplotype diversity (**Figure 6**). The dispersal corridors during the two periods are consistent, and both showed that the YGP area had a higher dispersal ratio, but the populations of the peripheral areas were rather isolated.

Correlation Between Genetic Diversity and Climatic Factors

The correlation analysis indicated that only the N_{LGM} was significantly correlated with GenPCoA1 ($P < 0.05$). A_r was associated with both longitude and latitude ($P < 0.05$). Cluster A (C_A) and GenPCoA1 were both associated with latitude ($P < 0.05$). However, H_e was not associated with N_{Pre}, $N_{stabLGM}$, N_{LGM}, latitude, nor longitude ($P > 0.05$) (**Table 3**).

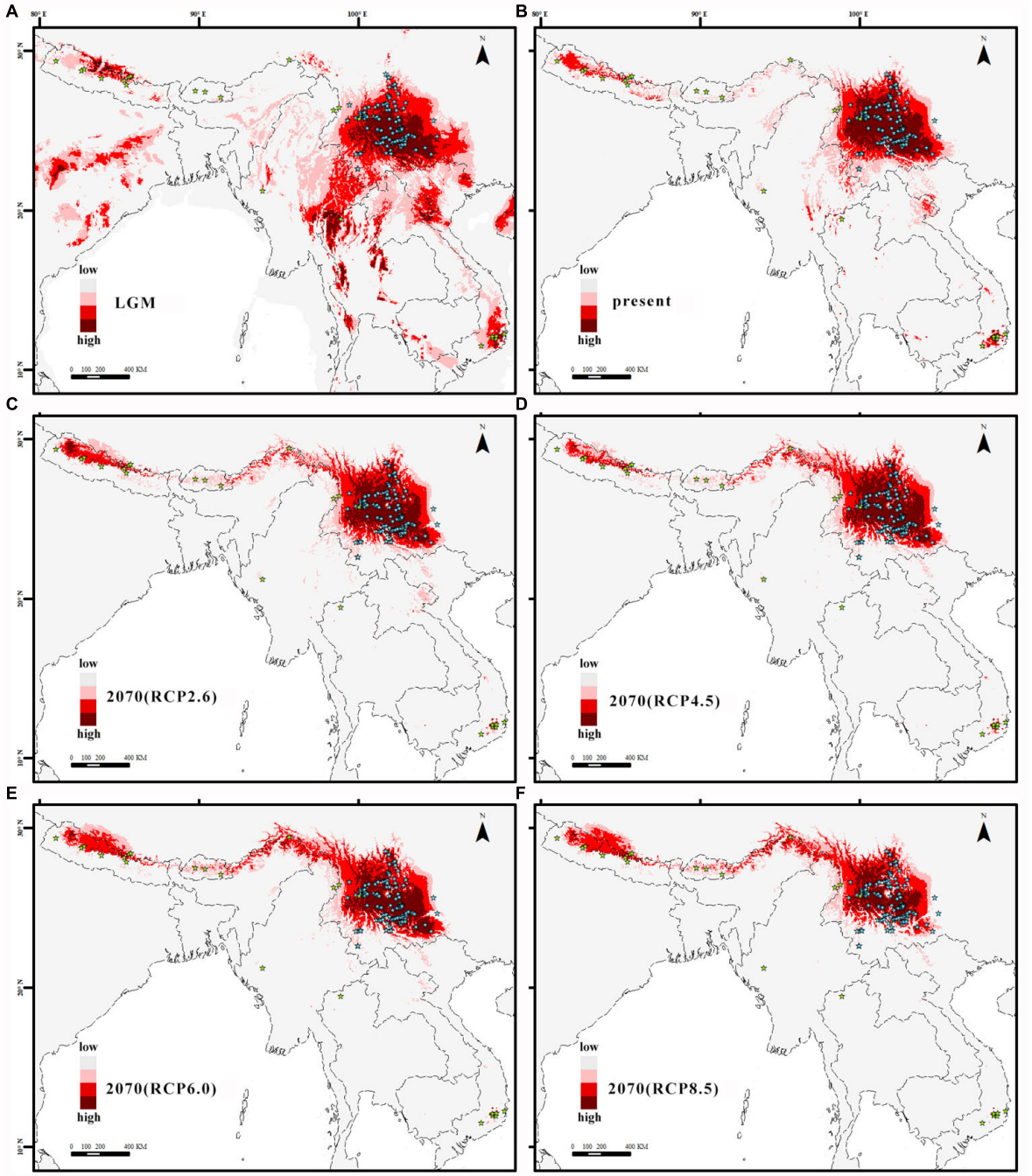


FIGURE 5 | (A) the last glacial maximum (LGM), **(B)** the present, **(C)** in 2070 under the RCP2.6 scenario, **(D)** in 2070 under the RCP4.5 scenario, **(E)** in 2070 under the RCP6.0 scenario, and **(F)** in 2070 under the RCP8.5 scenario. The color from pink to red represents the fitness zone, from low to high respectively.

DISCUSSION

Genetic Diversity Pattern of the *Quercus franchetii* Complex

Our study revealed a high genetic diversity estimate for the *Q. franchetii* complex (H_T, 0.982; A_r, 3.18–4.34), which is similar to the genetic diversity of other evergreen oaks in southwestern China, e.g., *Q. schottkyana* (H_T, 0.828; A_r, 4.83–7.78; Jiang et al., 2016), *Q. kerrii* (H_T, 0.71; A_r, 2.27–3.20; Jiang et al., 2018), *Q. delavayi* (H_T, 0.907; A_r, 3.750–5.237; Xu et al., 2020), and the *Q. cocciferoides* complex (H_T, 0.904; Liu, 2019). Comparatively, the genetic diversity for deciduous oaks seems lower, e.g., in the eight European white oaks (H_T, 0.635–0.847) (Petit et al., 2002), *Quercus variabilis* (H_T, 0.888, in 50 populations in East

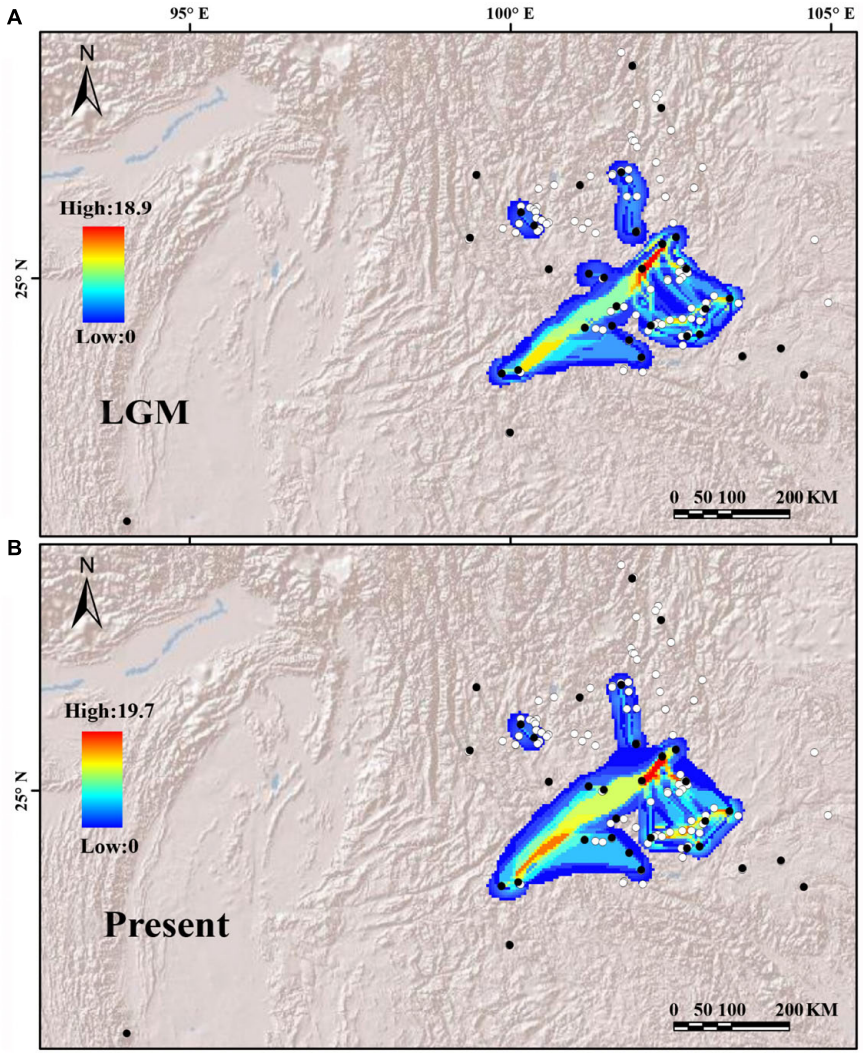


FIGURE 6 | Potential dispersal corridors of the *Quercus franchetii* complex during **(A)** the last glacial maximum (LGM) and **(B)** the present. Both white and black dots are the occurrence points of the *Q. franchetii* complex, with the black dots indicating sampling sites used in this study and white points indicating occurrence sites obtained from herbarium records. Colors in **(A,B)** from blue to red represent the potential of species dispersal corridors from low to high.

TABLE 3 | Correlation between genetic diversity, habitat and geographical factors of *Quercus franchetii* complex.

	Ar				*He*				C_A				GenPCoA1			
	Estimate	**Se^a**	**t**	***p***	**Estimate**	**Se^a**	**t**	***p***	**Estimate**	**Sea**	**t**	***p***	**Estimate**	**Sea**	**t**	***p***
(Intercept)	−7.44	4.00	−1.86	0.07	−0.70	0.59	−1.18	0.25	4.03	1.50	2.68	0.01*	14.42	6.35	2.27	0.03*
N_{Pre}	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
$N_{stadLGM}$	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
N_{LGM}	–	–	–	–	–	–	–	–	–	–	–	–	1.31	0.51	−2.98	0.01*
longitude	0.08	0.04	2.16	0.04*	0.01	0.01	1.89	0.07	–	–	–	–	−0.11	0.06	−1.75	0.09
latitude	0.13	0.04	3.20	0.004 *	0.01	0.01	1.50	0.15	−0.15	0.06	−2.44	0.02*	−0.17	0.06	−2.98	0.01*

**Significant correlation ($P < 0.05$), Se^a, standard error.*

Asia) (Chen et al., 2012), *Q. mongolica* var. *crispula* (H_T, 0.827, in Japan) (Okaura et al., 2007), *Q. acutissima* (H_T, 0.791, in Southeast China) (Zhang et al., 2015). The deciduous species are mainly Miocene-Pliocene derived young lineages, but the divergence time of the evergreen oaks are much older that dated to the Eocene-Miocene (Hipp et al., 2020). It is easy

to understand that oak taxa with long evolutionary histories and wide distribution can accumulate higher levels of genetic diversity than those "young" species. Additionally, the high environmental heterogeneity of southwestern China can buffer the climate extremes of the Quaternary episode (Zhang et al., 2010; Kai and Jiang, 2014; Jiang et al., 2018). As a result, the habitats of southwestern China have had long-term stability without significant regional extinction events or distribution range shifts to allow the genetic diversity of the species can be maintained. All these factors contributed to the high genetic diversity level found in evergreen oaks in southwestern China.

Comparing the genetic structure of the sympatric close related species can better illustrate the factors determining genetic diversity pattern. Notably, there was significant genetic differentiation among populations in the *Q. franchetii* complex, with F_{ST} estimates for *cp*DNA and nSSRs of 0.87 and 0.369, respectively, which is very similar to that found in *Q. delavayi*, as the *cp*DNA of the both species show significant phylogeographic structure, IBD pattern, and high differentiation among the populations (Xu et al., 2020). However, in two other sympatric/parapatric oaks, *Q. kerrii* (Jiang et al., 2018) and *Q. schottkyana* (Jiang et al., 2016), only low differentiation among the populations without phylogenetic structure were found (**Table 4**). The biological traits restrict gene flow and its efficiency, no doubt, can greatly impact the population genetic structure (Petit et al., 2003; Cavender-Bares et al., 2015; Xu et al., 2020). Pollen and seed mediated gene flow among the populations is different, as they have different dispersal efficiencies when barriers exist (e.g., rivers and high mountains). Generally, the seed-mediated gene flow among populations is more restricted in the species with instant germination seeds comparing to those species that seed with a period of dormancy. Consequently, the typical recalcitrant seed species shows significant phylogeographic structure in *cp*DNA makers. Vice versa, as pollen of oaks can disperse long distance, population differentiation revealed by biparental markers was much lower than that in maternal makers (Xu et al., 2020). However, the genetic structure of *Q. franchetii* seems not only determine by seed/pollen mediated gene flow efficiency. Although *Q. franchetii* has temporary seed dormancy (2–4 months; Xia et al., 2012), its seed size and tannin content similar to those of *Q. schottkyana*, but its population genetic structure is similar to that of *Q. delavayi*—a typical recalcitrant small seeds species. Thus, factors beyond seed germination schedule, seed size, and dispersal abilities also played important roles in shaping the population structure of these oaks in southwestern China, e.g., their ancestor distribution range and evolutionary history, etc.

Contemporary and historical factors shaped the genetic structure of organisms (Van Oppen et al., 2011; Hernawan et al., 2017; Li J. J. et al., 2017). Geological and climatic factors have been shown to influence the evolutionary histories of taxa and shape their genetic structures (Feng et al., 2014; Xing et al., 2014; Lu et al., 2018; Chen et al., 2020). Thus, the genetic structures of ancient lineages with long evolutionary histories and wide distribution range can essentially record more ancient geological events than in those of young lineages. The evergreen oak lineages in YGP and southwestern China, e.g., *Q. schottkayana* at 6.37 Ma (Jiang et al., 2016), *Q. kerrii* at 6–7 Ma (Jiang et al., 2018), and *Q. cocciferoides* at ca. 5 Ma (Jiang et al., 2019; Liu, 2019) are later derived "young" (originated at the late Neogene) and they generally show no (or very limited) IBD pattern among the populations. In contrasts, the early derived species *Q. delavayi* with crown node age at 10.92 Ma, and the *Q. franchetii* complex with crown node age at 30.7 Ma (95% HPD 16.7–43.9 Ma) had similar genetic structures and distinct IBD pattern. The different genetic structures detected in the "young" and "old" species might reflect the outcomes of past geological events on the biota at the different epochs, as the ancient geological events may imprint a genetic structure

TABLE 4 | Comparisons of the genetic diversity, genetic structure, and demographic dynamics of *Quercus franchetii* complex with *Q. schottkyana*, *Q. kerrii*, *Q. schottkyana,* and *Q. cocciferoides* complex.

Taxa	Genetic diversity		Genetic structure						Demographic change			IBD	
	H_T (se)	H_S (se)	G_{ST} (se)	N_{ST} (se)	Phylo structure	Network structure	F_{ST} (b)	F_{ST} (m)	*Na*	Mismatch	Neutral test	*R* (m)	*R* (b)
Q. franchetii complex	0.982 (0.01)	0.123 (0.04)	0.874 (0.04)	0.959* (0.02)	No	Star-like	0.369	0.87	0.42	No expansion	No expansion	0.052*	0.413*
Q. delavayi	0.907 (0.03)[a]	0.197 (0.05)[a]	0.782 (0.05)[a]	0.912* (0.03)[a]	Yes[a]	Geographic structure[a]	0.063[a]	0.938[a]	1.14[a]	No expansion[a]	No expansion[a]	0.587*[a]	0.365*[a]
Q. kerrii	0.71 (0.06)[b]	0.05 (0.02)[b]	0.93 (0.03)[b]	0.92 (0.04)[b]	No[b]	Star-like[b]	0.066[a]	0.894[a]	1.27[b]	Expansion[a]	No expansion[b]	–	–
Q. schottkyana	0.828 (0.06)[a]	0.341 (0.06)[a]	0.588 (0.07)[a]	0.615 (0.11)[a]	No[a]	Star-like[c]	0.075[a]	0.665[a]	1.03[c]	No expansion[c]	No expansion[c]	–	–
Q. cocciferoides Complex	0.904[d]	0.140[d]	0.845[d]	0.860[d]	Yes[d]	Geographic structure[d]	0.134[d]	0.946[d]	1.29[d]	No expansion[d]	No expansion[d]	0.312[d]	0.886[d]

*H_T, total haplotype diversity; H_S, within population diversity; G_{ST}, coefficient of genetic variation over all populations; N_{ST}, coefficient of genetic variation influenced by both haplotype frequencies and genetic distances between haplotypes; Phylo structure, phylogeographic structure; F_{ST}(b), population differentiation for nuclear; F_{ST}(m), population differentiation for maternally inherited; Na, habitat distribution area ratio; R(m), correlation of two matrices for maternally inherited; R(b), correlation of two matrices for nuclear; *N_{ST} differs from G_{ST} at $P < 0.05$.*

[a]Xu et al. (2020); [b]Jiang et al. (2018); [c]Jiang et al. (2016); [d]Articles to be published.

pattern in an "old" and widespread lineages, but not in those of the "young" lineages. However, our hypothesis requires further investigation, as the molecular markers we used in previous phylogeographical studies on oaks of southwestern China are not all the same. Regardless, these markers are informative to reveal the spatial genetic structures of these oaks, but it contains potential bias when comparing across the species to reveal the underlying mechanisms that shaped the genetic structures of these oaks. Further studies using universal high-throughput markers to scan oak populations in southwestern China and incorporating seed functional traits analyses are necessary to reveal the underlying drivers lead to the contemporary genetic structure found in these oaks.

The Impacts of Ancient Geological Events on the Divergence of Subtropical Lineages in the Southeastern Himalaya Fringe

The Oligocene was a key period in the formation of the modern topography of China. During this period, the high eastern and low western topographies of China were totally reversed (Ming, 2007; Di et al., 2018). The tectonic plates associated with the spreading of the East Asian marginal sea and the uplift of the eastern margin of the Tibetan Plateau exerted important influence on the topography and drainage of East Asia (Clark et al., 2004). This was accompanied by dramatic tectonically induced topographic changes and landscape development, which resulted in repeated river captures and reversals (Zheng, 2015). Paleomagnetic and stratigraphic evidence suggests that there was a wide Paleo-Red River drainage basin between the southeastern Tibetan Plateau and the South China Sea, which included Hoh-Xil, Songpan-Ganzi, northern Qiangtang, Yidun, and western Yangtze Terranes (**Figure 7**). Since the late Eocene, the hard collision between the Indian and Eurasian plates began. These massifs/blocks were extruded eastward under the resulting pressure (Chen et al., 2017). Meanwhile, the Ailao Shan-Red River fault zone began to slip rapidly. The Lanpin-Simao block began to rotate clockwise substantially, which blocked the upper and lower Yangtze River from continually flowing to the south along the paleo-drainage (**Figure 7**). As a result, the northern drainage basin of the paleo Red River disappeared, and the modern Red River began to become established (Replumaz and Tapponnier, 2003; Royden et al., 2008; Chen et al., 2017; He et al., 2021). These the Oligocene to early Miocene events greatly changed the regional biota, e.g., giving rise to the diversification of *Cautleya*, *Roscoea* (Zhao et al., 2016), *Badidae* (Rüber et al., 2004), and spiny frogs (*Paini*) (Che et al., 2010).

Our ancestral range reconstruction based on *cp*DNA data showed that the *Q. franchetii* complex once had a wide distribution in southwestern China and the southern Himalaya regions during the early Oligocene, followed by three vicariance events. Among these events, the NPR (Clade I) was first diversified during the mid-Oligocene (ca. 30 Ma), and then, during the late Oligocene and the early Miocene, the RR lineage and HDM lineage were derived, respectively. Within the main lineages, the fast divergence of the *cp*DNA haplotypes occurred during the late Miocene (**Figure 7A**). The NPR was located at the core area during early river capture; the hard collision between the Indian and Eurasian plates might have squeezed the plates in this region leading to river re-alignment, which induced the divergence of the NPR lineage (Clark et al., 2004; Yang et al., 2012; **Figure 7B**). During the late Oligocene, the Ailao Shan-Red River fault had an early left strike-slip that may have raised the barrier that blocked gene flow among the populations and promoted allopatric divergence (Hall et al., 2008; Fyhn et al., 2009; Lin et al., 2009; Deng et al., 2014). Thus, the RR lineage diverged (Clade II). During the early Miocene, large fault basins

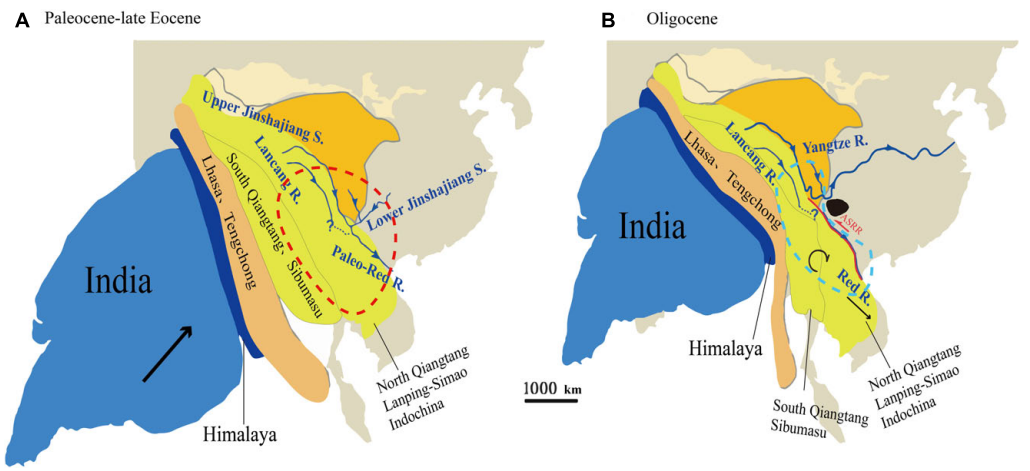


FIGURE 7 | Schematic geodynamic evolutionary model of the paleo-drainage evolution of East Asia and schematic diagram of the ancestral distribution of the *Quercus franchetii* complex from the Eocene to Oligocene (based on Replumaz and Tapponnier, 2003; Royden et al., 2008; Yan and Chen, 2018). **(A)** Paleocene to Late Eocene. The red dotted area indicates the possible ancestral distribution area of the *Q. franchetii* complex in the Paleogene; **(B)** the Oligocene. The black area indicates the area where the early evolution of the ancient Red River was captured (Nanpan River area). The blue dotted area shows the ancient Red River area in the Oligocene period.

were established corresponding to the fast uplift of the YGP and HDM regions (Zhang, 2012), which might dramatically change the regional topography and climate, eventually blocking gene flow in the two regions and leading to the divergence of the YGP lineage (Clade III) and HDM lineage (Clade IV). Followed by the late Neogene period fast HDM uplifts, the complex topography of southwestern China was eventually formed, which further restricted regional seed-mediated gene flow and promoted the divergence of *cp*DNA haplotypes. A similar scenario was also detected in other oaks, e.g., *Q. delavayi* (Xu et al., 2020), and *Q. aquifolioides* (Du et al., 2017), as well as plant lineages with wide distribution in semi-humid evergreen broadleaved forests on YGP region, e.g., *Primula secundiflora* (Wang et al., 2008), *Terminalia franchetii* (Zhang et al., 2011), and *Cycas multipinnata* (Gong et al., 2015). Such phenomenon suggested the regional biota might be impacted by similar environmental drivers.

In summary, the high biodiversity levels found in southwestern China are rooted deeply in the Oligocene. The early tectonic events during the Oligocene drove the main lineage splits, while the fast uplifts of the Himalayas during the Miocene-Pliocene increased environmental heterogeneity and established substantial dispersal barriers (Meng et al., 2017; Xing and Ree, 2017; Yuan et al., 2019). Then, the Quaternary climate fluctuation led to distribution range contractions and expansions of the species, as well as the occurrence of Asian winter monsoons, and the dry season in winter and spring in southwestern China further restricted gene flow between the core YGP region and its periphery (Su et al., 2013; Ye et al., 2019). All these geological and climatic factors interacted during different timespans to shape the contemporary divergence pattern of the biota of the southeastern Himalaya biodiversity hotspot. Our phylogeographic study indicated the Oligocene tectonic induced divergence in *Q. franchetii* complex, which is a supplement to the review of Renner (2016) concerning the Paleogene events contributing to the species richness of the East Himalayan biodiversity hotspot.

Moreover, the population genetic structures inferred from nSSR and *cp*DNA markers of *Q. franchet*ii complex were dissimilar. Notably, nSSRs mainly reflected the divergence between the populations in the core YGP region and the peripheral populations. The similar population genetic structure on nuclear genome was also reported in another sympatric species (*Q. cocciferoides*; Liu, 2019). All these evidences suggested the two species might underwent the similar selection pressure to trigger their divergence. The potential migration corridor analysis on the *Q. franchetii* complex showed that the populations in core YGP region maintained strong gene flows, but the marginal populations were mostly isolated since LGM (**Figure 6**). The rugged topography induced by the rapid YPG uplift during the late Neogene and (or) highly fragmented habitat of semi-humid evergreen forests in the peripheral areas around YGP might boost the allopatric divergence of these oaks. Nevertheless, another possibility of this pattern is that the quick evolution and possible backward evolution of SSR markers blurs the geological pattern. Further investigation using high throughput marker to illustrate the genetic structure at fine scale can provide a better understanding on the interplays between genetic diversity and environmental factors.

In contrast, the *cp*DNA data of *Q. franchetii* complex showed a much clearer phylogeographic structure, which has also been shown in other oaks, e.g., *Q. delavayi* (Xu et al., 2020), *Q. cocciferoides* (Liu, 2019), and *Q. aquifoliodes* (Du et al., 2017). Pollen- and seed-mediated gene flow have very different dispersal efficiencies in oaks (Du et al., 2017; Liu, 2019; Xu et al., 2020). In this study, nSSRs showed no significant differentiation of the populations in the East Red River, West Red River, and HDM regions, but *cp*DNA data showed a clear phylogeographic structure. This result suggested that these early tectonic activities during the Oligocene to early Miocene might have restricted seed-mediated gene flow in different regions, but only had limited impacts on pollen-induced gene flow.

In contrast, the niche modeling result suggested that the *Q. franchetii* complex populations are mainly located in the predicted suitable area. While the species complex experienced a southern contraction during the LGM, the distribution area at present and during the LGM largely overlapped in YGP. The quaternary glaciation had only minor impacts on its distribution. Likewise, the niche modeling results of *Osteomeles schwerinae* (Wang et al., 2015), *Q. schottkyana* (Jiang et al., 2016), and *Q. kerrii* (Jiang et al., 2018) from southwestern China show the similar pattern. Collectively, these studies suggest that central YGP region is an important refugia for species in semi-humid evergreen broadleaved forests in southwestern China.

CONCLUSION AND PERSPECTIVE

The spatial genetic structure is subject to environmental factors and the evolutionary process of the organism that affect genetic and genomic variation. The southeastern Himalaya fringe with extensive environmental changes since Cenozic high biodiversity. In our case, the population genetic diversity pattern of the *Q. franchetii* complex showed that the divergence of this subtropical lineage is rooted at the Oligocene. The tectonic events ever since this epoch might have restricted the regional seed-mediated gene flow, in turn triggered the early divergences of this subtropical woody lineage (Replumaz and Tapponnier, 2003; Clark et al., 2004; Zhang et al., 2011). Following, the rapid uplift-induced environmental heterogeneity in the Miocene in the southeastern Himalayas fringe, with subsequent Quaternary climatic fluctuations inducing distribution range expansions and contractions might further restrict the gene flow among the populations in core distribution and the peripheral areas (Huchon et al., 1994; Liu et al., 2006; Xu et al., 2010). These geological and climatic factors acted in a combined manner to boost the diversification of the subtropical biota in the southeastern Himalaya fringe. Our study provides an example that clearly reveals the evolutionary dynamics of the subtropical evergreen forests since the Oligocene in southwestern China for the first time, and demonstrated that except for the biological traits, the evolutionary history of the lineages are important factors impact the spatial genetic structures found in the evergreen oaks in YGP region. These results can provide important information on the formation of high biodiversity

level in southeast Himalayas, as well as conservation and safeguard this unique ecosystem on the background of global climate change.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, MW201294-MW201314, MW201315- MW201334, and MW201335- MW201352.

AUTHOR CONTRIBUTIONS

MD and X-LJ conceived and designed the experiments and were responsible for field collections and specimen identification. S-SZ performed the experiments. S-SZ and X-LJ analyzed the data. MD, S-SZ, and X-LJ wrote and revised the manuscript. Q-JH gave much advice on the details of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (31972858 and 31700174), the Southeast Asia Biodiversity Research Institute, Chinese Academy of Sciences (Y4ZK111B01), and the Shanghai Municipal Administration of Forestation and City Appearances (G182417).

ACKNOWLEDGMENTS

We would like to thank the editor and the two reviewers for their helpful comments and suggestions to improve the manuscript. We are grateful to Duo-Qing Lin for the help on DNA extraction, Yan-Shi Xiong for the help in the field collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.774232/full#supplementary-material

REFERENCES

Akhmetiev, M. A., and Zaporozhets, N. I. (2014). Paleogene events in Central Eurasia: their role in the flora and vegetation cover evolution, migration of phytochore boundaries, and climate changes. *Stratigr. Geol. Correlat.* 22, 312–335.

An, M., Deng, M., Zheng, S. S., and Song, Y. G. (2016). De novo transcriptome assembly and development of SSR markers of oaks *Quercus austrocochinchinensis* and *Q. kerrii* (Fagaceae). *Tree Genet. Genomes* 12:103.

Bandelt, H. J., Forster, P., and Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evolut.* 16, 37–48.

Bobrowski, M., and Schickhoff, U. (2017). Why input matters: Selection of climate data sets for modelling the potential distribution of a treeline species in the Himalayan region. *Ecol. Model.* 359, 92–102.

Bouchal, J. M., Mayda, S., Zetter, R., Grímsson, F., Akgün, F., and Denk, T. (2017). Miocene palynofloras of the Tınaz lignite mine, Muğla, southwest Anatolia: Taxonomy, palaeoecology and local vegetation change. *Rev. Palaeobot. Palynol.* 243, 1–36. doi: 10.1016/j.revpalbo.2017.02.010

Brown, J. L. (2014). SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *Methods Ecol. Evolut.* 5, 694–700. doi: 10.7717/peerj.4095

Cavender-Bares, J., Gonzalez-Rodriguez, A., Eaton, D. A. R., Hipp, A. L., Beulke, A., and Manos, P. S. (2015). Phylogeny and biogeography of the American live oaks (*Quercus* subsection *Virentes*): a genomic and population genetics approach. *Mol. Ecol.* 24, 3668–3687. doi: 10.1111/mec.13269

Chatterjee, S., Goswami, A., and Scotese, C. R. (2013). The longest voyage: Tectonic, magmatic, and paleoclimatic evolution of the Indian plate during its northward flight from Gondwana to Asia. *Gondwana Res.* 23, 238–267. doi: 10.1016/j.gr.2012.07.001

Che, J., Zhou, W. W., Hu, J. S., Yan, F., Papenfuss, T. J., Wake, D. B., et al. (2010). Spiny frogs (*Paini*) illuminate the history of the Himalayan region and Southeast Asia. *Proc. Natl. Acad. Sci.* 107, 13765–13770. doi: 10.1073/pnas.1008415107

Chen, D. M., Zhang, X. X., Kang, H. Z., Sun, X., Yin, S., Du, H. M., et al. (2012). Phylogeography of *Quercus variabilis* based on chloroplast DNA sequence in East Asia: multiple glacial refugia and mainland-migrated island populations. *PLoS One* 7:e47268. doi: 10.1371/journal.pone.0047268

Chen, J. H., Huang, Y., Brachi, B., Yun, Q. Z., Zhang, W., Lu, W., et al. (2019). Genome-wide analysis of Cushion willow provides insights into alpine plant divergence in a biodiversity hotspot. *Nat. Commun.* 10, 5230–5230. doi: 10.1038/s41467-019-13128-y

Chen, X. D., Yang, J., Feng, L., Zhou, T., Zhang, H., Li, H. M., et al. (2020). Phylogeography and population dynamics of an endemic oak (*Quercus fabri* Hance) in subtropical China revealed by molecular data and ecological niche modeling. *Tree Genet. Genomes* 16:2.

Chen, Y., Yan, M. D., Fang, X. M., Song, C. H., Zhang, W. L., Zan, J. B., et al. (2017). Detrital zircon U–Pb geochronological and sedimentological study of the Simao Basin, Yunnan: Implications for the Early Cenozoic evolution of the Red River. *Earth Planet. Sci. Lett.* 476, 22–33. doi: 10.1016/j.epsl.2017.07.025

Clark, M. K., Schoenbohm, L. M., Royden, L. H., Whipple, K. X., Burchfiel, B. C., Zhang, X., et al. (2004). Surface uplift, tectonics, and erosion of eastern Tibet from large-scale drainage patterns. *Tectonics* 23:TC1006.

Deng, J., Wang, Q. F., Li, G. J., and Santosh, M. (2014). Cenozoic tectono-magmatic and metallogenic processes in the Sanjiang region, southwestern China. *Earth Sci. Rev.* 138, 268–299. doi: 10.1016/j.earscirev.2014.05.015

Denk, T., Velitzelos, D., Güner, T. H., Bouchal, J. M., Grímsson, F., and Grimm, G. W. (2017). Taxonomy and palaeoecology of two widespread western Eurasian Neogene sclerophyllous oak species: *Quercus drymeja* Unger and *Q. mediterranea* Unger. *Rev. Palaeobot. Palynol.* 241, 98–128. doi: 10.1016/j.revpalbo.2017.01.005

Di, H. Z., Deng, B., Zhao, G. P., Ye, Y. H., and Qiu, J. W. (2018). The Evolution of the river system on the Yunnan-Guizhou plateau and formation process of the plateau based on modern river sediment. *Acta Geol. Sichuan* 38, 536–541.

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation precedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Du, F. K., Hou, M., Wang, W. T., Mao, K. S., and Hampe, A. (2017). Phylogeography of *Quercus aquifolioides* provides novel insights into the Neogene history of a major global hotspot of plant diversity in south-west China. *J. Biogeogr.* 44, 294–307.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294x.2005.02553.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Feng, X. Y., Wang, Y. H., and Gong, X. (2014). Genetic diversity, genetic structure and demographic history of *Cycas simplicipinna* (Cycadaceae) assessed by DNA sequences and SSR markers. *BMC Plant Biol.* 14:187. doi: 10.1186/1471-2229-14-187

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147, 915–925. doi: 10.1093/genetics/147.2.915

Fyhn, M. B. W., Boldreel, L. O., and Nielsen, L. H. (2009). Geological development of the Central and South Vietnamese margin: Implications for the establishment of the South China Sea, Indochinese escape tectonics and Cenozoic volcanism. *Tectonophysics* 478, 184–214. doi: 10.1016/j.tecto.2009.08.002

Gao, L. M., MÖller, M., Zhang, X. M., Hollingsworth, M. L., Liu, J., Mill, R. R., et al. (2007). High variation and strong phylogeographic pattern among *cp*DNA haplotypes in Taxus wallichiana (Taxaceae) in China and North Vietnam. *Mol. Ecol.* 16, 4684–4698. doi: 10.1111/j.1365-294X.2007.03537.x

Gao, H., Williamson, S., and Bustamante, C. D. (2007). A markov chain monte carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype. *Genetics* 176, 1635–1651. doi: 10.1534/genetics.107.072371

Gong, W., Chen, C., Dobeš, C., Fu, C. X., and Koch, M. A. (2008). Phylogeography of a living fossil: Pleistocene glaciations forced *Ginkgo biloba* L. (Ginkgoaceae) into two refuge areas in China with limited subsequent postglacial expansion. *Mol. Phylogenet. Evolut.* 48, 1094–1105. doi: 10.1016/j.ympev.2008.05.003

Gong, Y. Q., Zhan, Q. Q., Khang Sinh, N., Hiep Tien, N., Wang, Y. H., and Gong, X. (2015). The historical demography and genetic variation of the endangered *Cycas multipinnata* (Cycadaceae) in the Red River region, examined by chloroplast DNA sequences and microsatellite markers. *PLoS One* 10:e0117719. doi: 10.1371/journal.pone.0117719

Govaerts, R., and Frodin, D. G. (1998). *World Checklist and Bibliography of Fagales (Betulaceae, Corylaceae, Fagaceae and Ticodendraceae)*. London: Kew Publishing.

Grivet, D., Heinze, B., Vendramin, G. G., and Petit, R. J. (2001). Genome walking with consensus primers: application to the large single copy region of chloroplast DNA. *Mol. Ecol. Notes* 1, 345–349. doi: 10.1046/j.1471-8278.2001.00107.x

Guner, T. H., Bouchal, J. M., Kose, N., Goktas, F., Mayda, S., and Denk, T. (2017). Landscape heterogeneity in the Yatagan Basin (southwestern Turkey) during the middle Miocene inferred from plant macrofossils. *Palaeontogr. Abteilung B Palaeophytol. Palaeobot. Palaeophytol.* 296, 113–171. doi: 10.1127/palb/296/2017/113

Hall, R., Hattum, M. W. A. V., and Spakman, W. (2008). Impact of India–Asia collision on SE Asia: The record in Borneo. *Tectonophysics* 451, 0–389.

He, M. Y., Zheng, H. B., Clift, P. D., Bian, Z., Yang, Q., Zhang, B. H., et al. (2021). Paleogene Sedimentary Records of the Paleo-Jinshajiang (Upper Yangtze) in the Jianchuan Basin, Yunnan, SW China. *Geochem. Geophys. Geosyst.* 22:e2020GC009500.

Hernawan, U. E., Van Dijk, K. J., Kendrick, G. A., Feng, M., Biffin, E., Lavery, P. S., et al. (2017). Historical processes and contemporary ocean currents drive genetic structure in the seagrass Thalassia hemprichii in the Indo-Australian Archipelago. *Mol. Ecol.* 26, 1008–1021. doi: 10.1111/mec.13966

Hipp, A. L., Manos, P. S., Hahn, M., Avishai, M., Bodenes, C., Cavender-Bares, J., et al. (2020). Genomic landscape of the global oak phylogeny. *New Phytol.* 226, 1198–1212.

Huang, C. C., Chang, Y. T., and Bartholomew, B. (1999). "Fagaceae," in *Flora of China*, Vol. 4, eds C. Y. Wu and P. H. Raven (Beijing: Science Press and Missouri Botanical Garden Press), 380–400.

Huchon, P., Pichon, X. L., and Rangin, C. (1994). Indochina Peninsula and the collision of India and Eurasia. *Geology* 22, 27–30.

Hulce, D., Li, X., Snyder-Leiby, T., and Johathan Liu, C. S. (2011). GeneMarker® Genotyping Software: Tools to Increase the Statistical Power of DNA Fragment Analysis. *J. Biomol. Techniq. JBT* 22, S35–S36.

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Ji, Y. H., Liu, C. K., Landis, J. B., Deng, M., and Chen, J. H. (2020). Plastome phylogenomics of *Cephalotaxus* (Cephalotaxaceae) and allied genera. *Ann. Bot.* 127, 697–708. doi: 10.1093/aob/mcaa201

Jiang, X. L., An, M., Zheng, S. S., Deng, M., and Su, Z. H. (2018). Geographical isolation and environmental heterogeneity contribute to the spatial genetic patterns of *Quercus kerrii* (Fagaceae). *Heredity* 120, 219–233. doi: 10.1038/s41437-017-0012-7

Jiang, X. L., Deng, M., and Li, Y. (2016). Evolutionary history of subtropical evergreen broad-leaved forest in Yunnan Plateau and adjacent areas: an insight from *Quercus schottkyana* (Fagaceae). *Tree Genet. Genomes* 12:104.

Jiang, X. L., Hipp, A. L., Deng, M., Su, T., Zhou, Z. K., and Yan, M. X. (2019). East Asian origins of European holly oaks (*Quercus* section *Ilex* Loudon) via the Tibet-Himalaya. *J. Biogeogr.* 46, 2203–2214. doi: 10.1111/jbi.13654

Kai, H., and Jiang, X. L. (2014). Sky islands of southwest China I: an overview of phylogeographic patterns. *Chin. Sci. Bull.* 59, 585–597.

Kalinowski, S. T. (2005). HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. *Mol. Ecol. Notes* 5, 187–189.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., et al. (2017). Climatologies at high resolution for the Earth land surface areas. *Sci. Data* 4:170122.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evolut.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Li, J. G., Wu, Y. X., Batten, D. J., and Lin, M. Q. (2019). Vegetation and climate of the central and northern Qinghai–Xizang plateau from the Middle Jurassic to the end of the Paleogene inferred from palynology. *J. Asian Earth Sci.* 175, 35–48. doi: 10.1016/j.jseaes.2018.08.012

Li, S. H., Advokaat, E. L., Van Hinsbergen, D. J. J., Koymans, M., Deng, C. L., and Zhu, R. X. (2017). Paleomagnetic constraints on the Mesozoic-Cenozoic paleolatitudinal and rotational history of Indochina and South China: Review and updated kinematic reconstruction. *Earth Sci. Rev.* 171, 58–77.

Li, J. J., Hu, Z. M., Sun, Z. M., Yao, J. T., Liu, F. L., Fresia, P., et al. (2017). Historical isolation and contemporary gene flow drive population diversity of the brown alga Sargassum thunbergii along the coast of China. *BMC Evolution. Biol.* 17:246. doi: 10.1186/s12862-017-1089-6

Li, L., Abbott, R. J., Liu, B. B., Sun, Y. S., Li, L. L., Zou, J. B., et al. (2013). Pliocene intraspecific divergence and Plio-Pleistocene range expansions within *Picea likiangensis* (Lijiang spruce), a dominant forest tree of the Qinghai-Tibet Plateau. *Mol. Ecol.* 22, 5237–5255. doi: 10.1111/mec.12466

Lin, T. H., Lo, C. H., Chung, S. L., Hsu, F. J., Yeh, M. W., Lee, T. Y., et al. (2009). 40Ar/39Ar dating of the Jiali and Gaoligong shear zones: Implications for crustal deformation around the Eastern Himalayan Syntaxis. *J. Asian Earth Sci.* 34, 674–685.

Liu, F. Y., Wang, X. Q., Li, K., Sun, Y. Y., Zhang, Z. X., and Zhang, C. H. (2012). Species composition and diversity characteristics of *Quercus franchetii* communities in dry-hot valley of Jinsha River. *Guihaia* 32, 56–62.

Liu, F. Y., Zhang, Z. X., Wang, X. Q., Li, K., Sun, Y. Y., and Zhang, C. H. (2011). Effects of habitat heterogeneity on early growth of *Quercus franchetii* natural regeneration seedlings in the Jinsha river dry-hot valley. *Chin. J. Appl. Environ. Biol.* 17, 338–344. doi: 10.3724/sp.j.1145.2011.00338

Liu, J. Q., Gao, T. G., Chen, Z. D., and Lu, A. M. (2002). Molecular phylogeny and biogeography of the Qinghai-Tibet Plateau endemic *Nannoglottis* (Asteraceae). *Mol. Phylogenet. Evolut.* 23, 307–325. doi: 10.1016/s1055-7903(02)00039-8

Liu, J. Q., Wang, Y. J., Wang, A. L., Hideaki, O., and Abbott, R. J. (2006). Radiation and diversification within the *Ligularia-Cremanthodium-Parasenecio* complex (Asteraceae) triggered by uplift of the Qinghai-Tibetan Plateau. *Mol. Phylogenet. Evolut.* 38, 31–49. doi: 10.1016/j.ympev.2005.09.010

Liu, J., Möller, M., Provan, J., Gao, L. M., Poudel, R. C., and Li, D. Z. (2013). Geological and ecological factors drive cryptic speciation of yews in a biodiversity hotspot. *New Phytol.* 199, 1093–1108. doi: 10.1111/nph.12336

Liu, R. B. (2019). *Phylogeography of Quercus cocciferoides complex*. Shanghai: Shanghai Normal University.

Lu, L. M., Mao, L. F., Yang, T., Ye, J. F., Liu, B., Li, H. L., et al. (2018). Evolutionary history of the angiosperm flora of China. *Nature* 554:234.

Luikart, G., Sherwin, W. B., Steele, B. M., and Allendorf, F. W. (1998). Usefulness of molecular markers for detecting population bottlenecks via monitoring genetic change. *Mol. Ecol.* 7, 963–974. doi: 10.1046/j.1365-294x.1998.00414.x

Manni, F., Guérard, E., and Heyer, E. (2004). Geographic patterns of (genetic, morphologic, linguistic, etc.) variation : how barriers can be detected by "Monmonier's algorithm". *Hum. Biol.* 76, 173–190. doi: 10.1353/hub.2004.0034

Meng, H. H., Su, T., Gao, X. Y., Li, J., Jiang, X. L., Sun, H., et al. (2017). Warm–Cold colonization: Response of oaks to uplift of the Himalaya–Hengduan Mountains. *Mol. Ecol.* 26, 3276–3294. doi: 10.1111/mec.14092

Meng, Y., Wen, J., Nie, Z. L., Sun, H., and Yang, Y. P. (2008). Phylogeny and biogeographic diversification of *Maianthemum* (Ruscaceae: Polygonatae). *Mol. Phylogenet. Evolut.* 49, 424–434. doi: 10.1016/j.ympev.2008.07.017

Ming, Q. Z. (2007). A study on the neotectonic division & environment evolution of Qing-Zang plateau & three parallel rivers area. *Yunnan Geol.* 26, 387–396. doi: 10.1093/rpd/ncw067

Okaura, T., Quang, N. D., Ubukata, M., and Harada, K. (2007). Phylogeographic structure and late Quaternary population history of the Japanese oak *Quercus mongolica* var. *crispula* and related species revealed by chloroplast DNA variation. *Genes Genet. Syst.* 82, 465–477. doi: 10.1266/ggs.82.465

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460

Petit, R. J., Bodénès, C., Ducousso, A., Roussel, G., and Kremer, A. (2003). Hybridization as a mechanism of invasion in oaks. *New Phytol.* 161, 151–164. doi: 10.1046/j.1469-8137.2003.00944.x

Petit, R. J., Csaikl, U. M., Bordacs, S., Burg, K., Coart, E., Cottrell, J., et al. (2002). Chloroplast DNA variation in European white oaks: Phylogeography and patterns of diversity based on data from over 2600 populations. *For. Ecol. Manage.* 156:5.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecol. Model.* 190, 231–259.

Phillips, S. J., Dudík, M., and Schapire, R. (2004). "A maximum entropy approach to species distribution modeling," in *Proceedings of the twenty-first international conference on Machine learning*, (Alberta: ICML).

Pons, O., and Petit, R. J. (1996). Measwring and testing genetic differentiation with ordered versus unordered alleles. *Genetics* 144, 1237–1245. doi: 10.1093/genetics/144.3.1237

Posada, D., and Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14, 817–818. doi: 10.1093/bioinformatics/14.9.817

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

Qu, Y. H., Ericson, P. G. P., Quan, Q., Song, G., Zhang, R. Y., Gao, B., et al. (2014). Long-term isolation and stability explain high genetic diversity in the Eastern Himalaya. *Mol. Ecol.* 23, 705–720. doi: 10.1111/mec.12619

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Ree, R. H., and Smith, S. A. (2008). Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systemat. Biol.* 57, 4–14. doi: 10.1080/10635150701883881

Ree, R. H., Moore, B. R., Webb, C. O., and Donoghue, M. J. (2005). A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution* 59, 2299–2311.

Renner, S. S. (2016). Available data point to a 4-km-high Tibetan Plateau by 40 Ma, but 100 molecular-clock papers have linked supposed recent uplift to young node ages. *J. Biogeogr.* 43, 1479–1487.

Replumaz, A., and Tapponnier, P. (2003). Reconstruction of the deformed collision zone Between India and Asia by backward motion of lithospheric blocks. *J. Geophys. Res. Solid Earth* 108:2285.

Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* 8, 103–106. doi: 10.1111/j.1471-8286.2007.01931.x

Royden, L. H., Burchfiel, B. C., and Hilst, R. D. (2008). The geological evolution of the Tibetan Plateau. *Science* 321, 1054–1058. doi: 10.1126/science.1155371

Rozas, J., Ferrer-Mata, A., Sánchez-Delbarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evolut.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Rüber, L., Britz, R., Kullander, S. O., and Zardoya, R. (2004). Evolutionary and biogeographic patterns of the Badidae (Teleostei: Perciformes) inferred from mitochondrial and nuclear DNA sequence data. *Mol. Phylogenet. Evolut.* 32, 1010–1022. doi: 10.1016/j.ympev.2004.04.020

Shao, C. C., Shen, T. T., Jin, W. T., Mao, H. J., Ran, J. H., and Wang, X. Q. (2019). Phylotranscriptomics resolves interspecific relationships and indicates multiple historical out-of-North America dispersals through the Bering Land Bridge for the genus *Picea* (Pinaceae). *Mol. Phylogenet. Evolut.* 141:106610. doi: 10.1016/j.ympev.2019.106610

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., et al. (2005). The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot.* 92, 142–166. doi: 10.3732/ajb.92.1.142

Shen, L., Chen, X. Y., Zhang, X., Li, Y. Y., Fu, C. X., and Qiu, Y. X. (2005). Genetic variation of *Ginkgo biloba* L. (Ginkgoaceae) based on *cp*DNA PCR-RFLPs: inference of glacial refugia. *Heredity* 94, 396–401. doi: 10.1038/sj.hdy.6800616

Su, T., Farnsworth, A., Spicer, R. A., Huang, J., Wu, F. X., Liu, J., et al. (2019). No high Tibetan Plateau until the Neogene. *Sci. Adv.* 5:eaav2189.

Su, T., Liu, Y. S., Jacques, F. M. B., Huang, Y. J., Xing, Y. W., and Zhou, Z. K. (2013). The intensification of the East Asian winter monsoon contributed to the disappearance of *Cedrus* (Pinaceae) in southwestern China. *Quaternary Res.* 80, 316–325. doi: 10.1016/j.yqres.2013.07.001

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolut.* 4:vey016. doi: 10.1093/ve/vey016

Taberlet, P., Gielly, L., Pautou, G., and Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol. Biol.* 17, 1105–1109. doi: 10.1007/bf00037152

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi: 10.1093/genetics/123.3.585

Van Oosterhout, C., Hutchinson, W. F., Wills, D. P. M., and Shipley, P. (2004). Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* 4, 535–538. doi: 10.1111/j.1471-8286.2004.00684.x

Van Oppen, M. J. H., Peplow, L. M., Kininmonth, S., and Berkelmans, R. A. Y. (2011). Historical and contemporary factors shape the population genetic structure of the broadcast spawning coral, Acropora millepora, on the Great Barrier Reef. *Mol. Ecol.* 20, 4899–4914. doi: 10.1111/j.1365-294X.2011.05328.x

Vandergast, A. G., Perry, W. M., Lugo, R. V., and Hathaway, S. A. (2011). Genetic landscapes GIS Toolbox: tools to map patterns of genetic divergence and diversity. *Mol. Ecol. Resour.* 11, 158–161. doi: 10.1111/j.1755-0998.2010.02904.x

Wang, C. B., Wang, T., and Su, Y. J. (2014). Phylogeography of *Cephalotaxus oliveri* (Cephalotaxaceae) in relation to habitat heterogeneity, physical barriers and the uplift of the Yungui Plateau. *Mol. Phylogenet. Evolut.* 80, 205–216. doi: 10.1016/j.ympev.2014.08.015

Wang, F. Y., Gong, X., Hu, C. M., and Hao, G. (2008). Phylogeography of an alpine species *Primula secundiflora* inferred from the chloroplast DNA sequence variation. *J. Systemat. Evolut.* 46, 13–22.

Wang, J., Gao, P. X., Kang, M., Lowe, A. J., and Huang, H. W. (2009). Refugia within refugia: the case study of a canopy tree (*Eurycorymbus cavaleriei*) in subtropical China. *J. Biogeogr.* 36, 2156–2164.

Wang, Y. J., Liu, J. Q., and Miehe, G. (2007). Phylogenetic origins of the Himalayan endemic *Dolomiaea*, *Diplazoptilon* and *Xanthopappus* (Asteraceae: Cardueae) based on three DNA regions. *Ann. Bot.* 99, 311–322. doi: 10.1093/aob/mcl259

Wang, Z. W., Chen, S. T., Nie, Z. L., Zhang, J. W., Zhou, Z., Deng, T., et al. (2015). Climatic factors drive population divergence and demography: Insights based on the phylogeography of a riparian plant species endemic to the Hengduan Mountains and adjacent regions. *PLoS One* 10:e0145014. doi: 10.1371/journal.pone.0145014

Wangda, P., and Ohsawa, M. (2006). Structure and regeneration dynamics of dominant tree species along altitudinal gradient in a dry valley slopes of the Bhutan Himalaya. *For. Ecol. Manage.* 230, 136–150. doi: 10.1016/j.foreco.2006.04.027

Warren, D. L., Glor, R. E., and Turelli, M. (2010). ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography* 33, 607–611.

Xia, K., Daws, M. I., Stuppy, W., Zhou, Z.-K., and Pritchard, H. W. (2012). Rates of water loss and uptake in recalcitrant fruits of *Quercus* species are determined by pericarp anatomy. *PLoS One* 7:e47368. doi: 10.1371/journal.pone.0047368

Xing, Y. W., and Ree, R. H. (2017). Uplift-driven diversification in the Hengduan Mountains, a temperate biodiversity hotspot. *Proc. Natl. Acad. Sci.* 114, E3444–E3451. doi: 10.1073/pnas.1616063114

Xing, Y. W., Onstein, R. E., Carter, R. J., Stadler, T., and Peter Linder, H. (2014). Fossils and a large molecular phylogeny show that the evolution of species richness, generic diversity, and turnover rates are disconnected. *Evolution* 68, 2821–2832. doi: 10.1111/evo.12489

Xu, J., Deng, M., Jiang, X. L., Westwood, M., Song, Y. G., and Turkington, R. (2015). Phylogeography of *Quercus glauca* (Fagaceae), a dominant tree of East Asian subtropical evergreen forests, based on three chloroplast DNA interspace sequences. *Tree Genet. Genomes* 11:805.

Xu, J., Song, Y. G., Deng, M., Jiang, X. L., Zheng, S. S., and Li, Y. (2020). Seed germination schedule and environmental context shaped the population genetic structure of subtropical evergreen oaks on the Yun-Gui Plateau, Southwest China. *Heredity* 124, 499–513. doi: 10.1038/s41437-019-0283-2

Xu, T. T., Abbott, R. J., Milne, R. I., Mao, K., Du, F. K., Wu, G. L., et al. (2010). Phylogeography and allopatric divergence of cypress species (*Cupressus* L.) in the Qinghai-Tibetan Plateau and adjacent regions. *BMC Evolution. Biol.* 10:194. doi: 10.1186/1471-2148-10-194

Yan, M. D., and Chen, Y. (2018). Detrital zircon U-Pb age analyses of the Early Cenozoic sediments from the Simao Basin and evolution of the paleo-Red River drainage system. *Quat. Sci.* 38, 130–144. doi: 10.11928/j.issn.1001-7410.2018.01.11

Yang, J., Yang, J. X., and Chen, X. Y. (2012). A re-examination of the molecular phylogeny and biogeography of the genus *Schizothorax* (Teleostei: Cyprinidae) through enhanced sampling, with emphasis on the species in the Yunnan–Guizhou Plateau, China. *J. Zool. Syst. Evolution. Res.* 50, 184–191. doi: 10.1111/j.1439-0469.2012.00661.x

Ye, X. Y., Ma, P. F., Yang, G. Q., Guo, C., Zhang, Y. X., Chen, Y. M., et al. (2019). Rapid diversification of alpine bamboos associated with the uplift of the Hengduan Mountains. *J. Biogeogr.* 46, 2678–2689. doi: 10.1111/jbi.13723

Yu, Y., Harris, A. J., Blair, C., and He, X. J. (2015). RASP (Reconstruct Ancestral State in Phylogenies): A tool for historical biogeography. *Mol. Phylogenet. Evolut.* 87, 46–49. doi: 10.1016/j.ympev.2015.03.008

Yuan, Z., Jiang, J. B., Dong, Y., Zhao, Q., Gao, X., and Qiu, J. P. (2019). The dispersal and diversification of earthworms (Annelida: Oligochaeta) related to paleogeographical events in the Hengduan Mountains. *Eur. J. Soil Biol.* 94:103118.

Zhang, L. S. (2012). *Palaeogeography of China: the Formation of China's Natural Environment*. Beijing: Science Press.

Zhang, M. W., Rao, D. Q., Yang, J. X., Yu, G. H., and Wilkinson, J. A. (2010). Molecular phylogeography and population structure of a mid-elevation montane frog *Leptobrachium ailaonicum* in a fragmented habitat of southwest China. *Mol. Phylogenet. Evolut.* 54, 47–58. doi: 10.1016/j.ympev.2009.10.019

Zhang, T. C., Comes, H. P., and Sun, H. (2011). Chloroplast phylogeography of *Terminalia franchetii* (Combretaceae) from the eastern Sino-Himalayan region and its correlation with historical river capture events. *Mol. Phylogenet. Evolut.* 60, 1–12. doi: 10.1016/j.ympev.2011.04.009

Zhang, X. W., Li, Y., Liu, C. Y., Xia, T., Zhang, Q., and Fang, Y. M. (2015). Phylogeography of the temperate tree species *Quercus acutissima* in China: Inferences from chloroplast DNA variations. *Biochem. Syst. Ecol.* 63, 190–197. doi: 10.1016/j.bse.2015.10.010

Zhao, J. L., Xia, Y. M., Cannon, C. H., Kress, W. J., and Li, Q. J. (2016). Evolutionary diversification of alpine ginger reflects the early uplift of the Himalayan–Tibetan Plateau and rapid extrusion of Indochina. *Gondwana Res.* 32, 232–241. doi: 10.1016/j.gr.2015.02.004

Zheng, H. B. (2015). Birth of the Yangtze River: age and tectonic-geomorphic implications. *Natl. Sci. Rev.* 2, 438–453. doi: 10.1093/nsr/nwv063

Zheng, S. S. (2021). *Phylogeography of Quercus franchetii complex (Fagaceae)*. Shanghai: Shanghai Institute of Technology.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 24 November 2021
doi: 10.3389/fpls.2021.767478

Sage Insights Into the Phylogeny of *Salvia*: Dealing With Sources of Discordance Within and Across Genomes

Jeffrey P. Rose[1,2], Ricardo Kriebel[2], Larissa Kahan[2], Alexa DiNicola[2], Jesús G. González-Gallegos[3], Ferhat Celep[4], Emily M. Lemmon[5], Alan R. Lemmon[6], Kenneth J. Sytsma[2] and Bryan T. Drew[1]*

[1] Department of Biology, University of Nebraska at Kearney, Kearney, NE, United States, [2] Department of Botany, University of Wisconsin–Madison, Madison, WI, United States, [3] CONACYT, Instituto Politeicnico Nacional, CIIDIR – Durango, Durango, Mexico, [4] Department of Biology, Faculty of Arts and Sciences, Kırıkkale University, Yahşihan, Turkey, [5] Department of Biological Science, Florida State University, Tallahassee, FL, United States, [6] Department of Scientific Computing, Florida State University, Tallahassee, FL, United States

OPEN ACCESS

Edited by:
Stefan Wanke, Dresden University of Technology, Germany

Reviewed by:
Itzi Fragoso-Martínez, Instituto de Ecología (INECOL), Mexico
Roser Vilatersana, Consejo Superior de Investigaciones Científicas, Spanish National Research Council (CSIC), Spain

***Correspondence:**
Jeffrey P. Rose
rosej@unk.edu

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *30 August 2021*
Accepted: *22 October 2021*
Published: *24 November 2021*

Citation:
Rose JP, Kriebel R, Kahan L, DiNicola A, González-Gallegos JG, Celep F, Lemmon EM, Lemmon AR, Sytsma KJ and Drew BT (2021) Sage Insights Into the Phylogeny of Salvia: Dealing With Sources of Discordance Within and Across Genomes. Front. Plant Sci. 12:767478. doi: 10.3389/fpls.2021.767478

Next-generation sequencing technologies have facilitated new phylogenomic approaches to help clarify previously intractable relationships while simultaneously highlighting the pervasive nature of incongruence within and among genomes that can complicate definitive taxonomic conclusions. *Salvia* L., with ∼1,000 species, makes up nearly 15% of the species diversity in the mint family and has attracted great interest from biologists across subdisciplines. Despite the great progress that has been achieved in discerning the placement of *Salvia* within Lamiaceae and in clarifying its infrageneric relationships through plastid, nuclear ribosomal, and nuclear single-copy genes, the incomplete resolution has left open major questions regarding the phylogenetic relationships among and within the subgenera, as well as to what extent the infrageneric relationships differ across genomes. We expanded a previously published anchored hybrid enrichment dataset of 35 exemplars of *Salvia* to 179 terminals. We also reconstructed nearly complete plastomes for these samples from off-target reads. We used these data to examine the concordance and discordance among the nuclear loci and between the nuclear and plastid genomes in detail, elucidating both broad-scale and species-level relationships within *Salvia*. We found that despite the widespread gene tree discordance, nuclear phylogenies reconstructed using concatenated, coalescent, and network-based approaches recover a common backbone topology. Moreover, all subgenera, except for *Audibertia*, are strongly supported as monophyletic in all analyses. The plastome genealogy is largely resolved and is congruent with the nuclear backbone. However, multiple analyses suggest that incomplete lineage sorting does not fully explain the gene tree discordance. Instead, horizontal gene flow has been important in both the deep and more recent history of

Salvia. Our results provide a robust species tree of *Salvia* across phylogenetic scales and genomes. Future comparative analyses in the genus will need to account for the impacts of hybridization/introgression and incomplete lineage sorting in topology and divergence time estimation.

Keywords: anchored hybrid enrichment, cyto-nuclear discordance, distance metrics, incongruence, Lamiaceae, Robinson–Foulds distance, *Salvia*

INTRODUCTION

It has long been recognized that when generating multilocus nucleotide sequence data, different datasets may generate alternative gene tree topologies and, by extension, differing hypotheses of relationships among species (Pamilo and Nei, 1988; Rieseberg and Soltis, 1991; Maddison, 1997). The underlying causes for why such differing gene tree topologies may exist (apart from analytical artifacts) have been well-discussed in the literature, and include gene duplication, incomplete lineage sorting (ILS), lateral gene transfer, and introgression/hybridization (Degnan, 2018). These processes are not mutually exclusive, and the history of one locus may be shaped by multiple processes. For many years, the solution to deal with such discordance was to analyze incongruent datasets separately, attempt to reconcile these topologies into a consensus tree, or concatenate all loci together to generate a "total-evidence" hypothesis of the species relationships (Ané et al., 2007). The concept of genomic concordance, coupled with new methods for estimating species trees while taking into account ILS and/or horizontal gene flow have been important advances in the field of systematic biology (Ané et al., 2007; Baum, 2007; Heled and Drummond, 2009; Mirarab et al., 2014; Yu and Nakhleh, 2015; Edwards et al., 2016; Solís-Lemus and Ané, 2016).

Contemporaneous with these computational advances, new sequencing technologies have facilitated relatively easy and cost-effective sequencing of complete organellar genomes and hundreds to thousands of nuclear loci. The confluence of these two areas of biology has made for an exciting time for studies in systematic biology but has also presented challenges as to how to best analyze these datasets. For example, individual loci may have relatively little phylogenetic information and thus, confound analyses that rely on individual gene trees. In addition, the computational ability of many current algorithms are challenged by the number of terminals present in the species tree and especially the phylogenetic network that a researcher wishes to estimate (Hejase and Liu, 2016; Solís-Lemus and Ané, 2016; Rose et al., 2021). Despite these challenges, an ever-increasing proportion of phylogenomic studies employ methods that account for sources of intra-genomic discordance, especially due to ILS. While methods that employ the multispecies coalescent only are relatively fast and tractable on datasets with dozens to hundreds of terminals, it is increasingly clear that hybridization and introgression are important processes at both shallow and deep phylogenetic scales, and this affects all branches of the Tree of Life (Folk et al., 2018). If horizontal gene flow has been operative, the species tree estimated by methods that only account for ILS may differ substantially from the "true" species tree not only topologically, but also in branch lengths (Leaché et al., 2014). The misestimation of both properties may impact myriad downstream analyses.

Apart from discordance among nuclear loci, gene trees may differ among genomes. This phenomenon is well-known and often referred to as "cytonuclear discordance" (Rieseberg and Soltis, 1991). In plants, this is best demonstrated in cases of putative "chloroplast capture" which have been documented for decades (e.g., Smith and Sytsma, 1990). Such discordance has generally been taken as evidence of horizontal gene flow, even though organellar genomes are also susceptible to ILS, albeit with a much faster expected time to coalescence, relative to nuclear loci. Simulation studies have generally confirmed that most cases of chloroplast (technically plastid) capture are indeed best explained by horizontal gene flow, rather than ILS (Folk et al., 2017; Morales-Briones et al., 2018; Rose et al., 2021).

Therefore, a better understanding of the evolutionary history of clades requires an assessment of the contribution of each of the multiple processes responsible for the discordance among loci. This assessment is important not only for producing a robust phylogenetic hypothesis, but also for selecting methods, taxa, and loci appropriate for the downstream analyses of trait evolution, historical biogeography, and diversification rates. Robust phylogenetic hypotheses are also crucial for making informed decisions to ensure an accurate and stable taxonomic circumscription, from the species level to higher-level classifications.

Sage and its relatives (*Salvia* L.) comprise ∼1,000 species, with a subcosmopolitan distribution across a diversity of habitat types (Kriebel et al., 2019). It is the largest genus within the mint family (Lamiaceae) and one of the largest genera of plants. There are three broadly defined centers of diversity of *Salvia* (Walker et al., 2004): East Asia (∼100 spp.; subgenus (subg.) *Glutinaria*; Hu et al., 2018), the Mediterranean (∼250 spp.; subg. *Salvia*, *Sclarea*), and especially, Mexico, Central, and South America (∼580 spp., subg. *Calosphace*; González-Gallegos et al., 2020). *Salvia* is not only of interest from an economic perspective, given its culinary use (e.g., chia: *S. hispanica* L.; rosemary: *S. rosmarinus* (L.) Spenn.; sage: *S. officinalis* L.), but also in its horticultural importance (e.g., blue sage: *S. nemorosa* L.; pineapple sage: *S. elegans* Vahl; Russian sage*: S. yangii* B.T.Drew).

Salvia is florally diverse (Kriebel et al., 2019, 2020) and easily characterized by the presence of two stamens with an elongate (or swollen) anther connective, in addition to several micromorphological synapomorphies (Drew et al., 2017). In many species of *Salvia*, the connective has been variously modified – possibly multiple times – into a staminal lever mechanism to facilitate effective pollination

(Claßen-Bockhoff et al., 2003, 2004; Walker and Sytsma, 2007; Wester and Claßen-Bockhoff, 2007; Celep et al., 2020).

As a result of its practical importance to humans, distribution, size and taxonomic complexity, and unique pollination biology, *Salvia* has received considerable attention from systematists and pollination biologists. Early in the study of the phylogenetic placement of *Salvia*, it was realized that the genus was polyphyletic or broadly paraphyletic, with several smaller genera embedded within it (Walker and Sytsma, 2007). Subsequent phylogenetic analyses have confirmed that five previously recognized small genera (*Dorystaechas* Boiss. & Heldr., *Meriandra* Benth., *Perovskia* Kar., *Rosmarinus* L., and *Zhumeria* Rech.f. & Wendelbo) are nested within several clades of *Salvia* (Walker and Sytsma, 2007; Drew and Sytsma, 2012; Will and Claßen-Bockhoff, 2014, 2017; Drew et al., 2017). To accommodate these small genera, Drew et al. (2017) and Kriebel et al. (2019) presented an expanded concept of *Salvia*, recognizing a total of 11 subgenera, although their informal circumscription of subg. "*Heterosphace*" represents a geographically diverse assemblage of lineages.

To date, most phylogenetic studies of *Salvia* have relied on plastid or nuclear ribosomal external transcribed spacer (ETS) and especially internal transcribed spacer (ITS) sequences (Walker and Sytsma, 2007; Jenks et al., 2013; Will and Claßen-Bockhoff, 2014, 2017; Dizkirici et al., 2015; Walker et al., 2015; Fragoso-Martínez et al., 2018; Hu et al., 2018). The resolution at multiple phylogenetic scales with these markers is variable by clade and, while there has been evidence for several deeper-level clades along the backbone of *Salvia*, relationships among them have usually either not been resolved or well supported. Discordance among plastid and nuclear ribosomal loci is generally found but not well-discussed or quantified (but see Walker et al., 2015). In cases where relationships differ across studies and marker sets, it is not clear if the differences are due to a true discordance in genealogical history or are from errors in the phylogenetic estimation (cf. Will and Claßen-Bockhoff, 2017: Figure 1). Drew et al. (2017) further investigated the backbone relationships in *Salvia* using two low-copy nuclear loci. While several key nodes remained unresolved and there was a clear conflict between the loci, they found an increased resolution for the backbone relationships. More recently, Zhao et al. (2020) used complete plastomes from seven of 11 subgenera and recovered a nearly fully resolved backbone across *Salvia* except for uncertainty in the placements of subg. *Perovskia* and *Rosmarinus*.

Multilocus nuclear datasets from anchored hybrid enrichment (AHE) have been successfully used to resolve deep and shallow level relationships across multiple angiosperm lineages, including *Salvia* (Fragoso-Martínez et al., 2017; Kriebel et al., 2019). Previously, we presented a species tree of 35 *Salvia* exemplars from 10 of 11 subgenera based on 316 nuclear loci using concatenation and one coalescent method (Kriebel et al., 2019). While this topology is congruent with the plastome phylogeny of Zhao et al. (2020) in areas where the two studies overlap in subgeneric sampling, several factors bear further consideration in Kriebel et al. (2019). First, the branching order of subg. "*Heterosphace*", *Salvia*, and *Sclarea* is not fully supported. Second, the monophyly of subg. *Audibertia* is not fully supported. Third, within subg. *Calosphace*, section (sect.) *Axillares* was recovered as a sister to the "*Hastatae* clade". instead of sister to all other *Calosphace*, in conflict with several previous studies (Jenks et al., 2013; Drew et al., 2017; Fragoso-Martínez et al., 2018). The placement of sect. *Axillares* has important implications for understanding character evolution in subg. *Calosphace* (e.g., Fragoso-Martínez et al., 2018; Kriebel et al., 2019, 2020, 2021).

Given the relatively sparse sampling of *Salvia* diversity from previous phylogenomic analyses, as well as the limited exploration of any discordance surrounding the backbone relationships in *Salvia*, we aimed to sample the species diversity in *Salvia* better using AHE to fulfill several goals. (1) Fully resolve the backbone of *Salvia* and assess the monophyly of the subgenera, quantifying discordance and accounting for both ILS and horizontal gene flow. (2) Generate a species tree for a much broader species-level sampling of *Salvia*, testing the efficacy of the AHE data for resolving shallow-level relationships. (3) Examine cytonuclear discordance at multiple phylogenetic scales by mining off-target organellar reads.

MATERIALS AND METHODS

Species Sampling in *Salvia* and Outgroups

In total, the analyses consisted of 190 samples, including 179 *Salvia* and all subgenera recognized by Kriebel et al. (2019) except for the small subg. *Meriandra* (Benth.) J.B.Walker, B.T.Drew & J.G.González. For ease of discussion, we will refer to the "*Heterosphace*" clade as a subgenus, although we acknowledge that this is not a formally named taxon. Within the subtribe Salviinae, we sampled the clade sister to *Salvia* (six species of *Lepechinia* and one species of *Melissa*: Drew and Sytsma, 2011, 2012, 2013). The ultimate outgroup consisted of samples from all the remaining subtribes of Mentheae, which represents a monophyletic group (Drew and Sytsma, 2012): Lycopinae (*Lycopus uniflorus* Michx.), Menthinae (*Clinopodium mexicanum* (Benth.) Govaerts), Nepetinae (*Agastache pallida* (Lindl.) Cory), and Prunellinae (*Prunella vulgaris* L.).

Anchored Hybrid Enrichment: Library Preparation, Enrichment, Sequencing, and Nuclear Locus Assembly

Total DNA was extracted from the silica gel-dried or fresh leaf tissue using a DNeasy Plant Mini Kit (Qiagen, Valencia, CA, United States). The DNA concentrations were verified using a Qubit® 2.0 Fluorometer (Life Technologies, Eugene, OR, United States). We used the AHE method (Lemmon and Lemmon, 2012). As the samples were sequenced across several years of study, we enriched them using slightly different probe sets (**Supplementary Data Sheet S1**). The samples sequenced early on in our studies utilized a generic angiosperm kit that targets 517 loci (Buddenhagen et al., 2016; Mitchell et al., 2017) or the *Salvia*-specific probes utilized in Kriebel et al. (2019), designed using the genome skimming of several *Salvia*

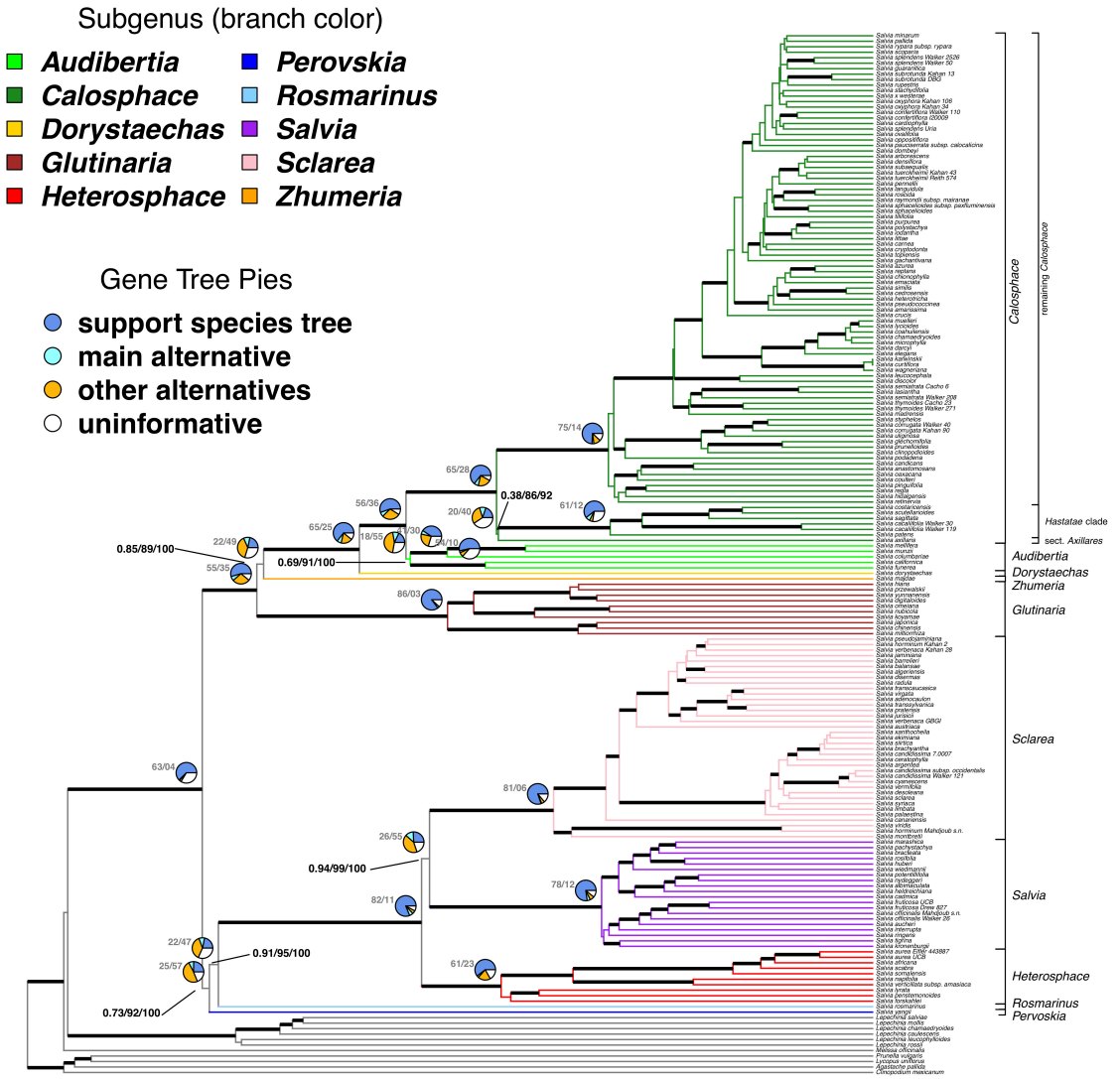


FIGURE 1 | The ASTRAL species tree of *Salvia* and outgroups. The ingroup branches are colored by subgenus and the subgenera are also labeled to the right. The major clades as discussed in the text are also indicated for subg. *Calosphace*. Thickened branches denote those with ≥0.95 ASTRAL local posterior probability. Pies at major nodes summarize the percentage of various phylogenetic signals across 101 gene trees which can be rooted. The numbers at the left of the pies show the total number of gene trees in which the clade is found, followed by the total number of gene trees that conflict with that clade. The remainder of the gene trees, if any, do not provide information on that particular relationship. The numbers at selected, incompletely supported nodes show the ASTRAL local posterior probability followed by the ASTRAL bootstrap support and the bootstrap support from the concatenated maximum likelihood analysis, summarized on the ASTRAL species tree. For support across all branches, see **Supplementary Figures S1, S2, S4**. For support on the best-scoring maximum likelihood tree, see **Supplementary Figure S3**.

species. The *Salvia*-specific probes targeted the same regions as the generic angiosperm kit, only neighboring exons could be combined and target regions extended, resulting in a target of 291 moderately conserved, low copy nuclear loci and their variable flanks. The library preparation, enrichment, assembly, and alignment of nuclear loci were performed at the Florida State University Center for Anchored Phylogenomics[1] and are described in detail in Kriebel et al. (2019). Because of the large number of samples, a substantial number of loci were lost during the orthology assessment, resulting in 123 recovered loci. Nine additional loci were lost during trimming and masking due to excessive missing data, resulting in a final dataset of 114 loci.

Nuclear Dataset 1: Complete Dataset

To examine the monophyly of the subgenera (when represented by multiple samples) and assess shallow-level relationships, we assembled species trees using all accessions. First, we concatenated all loci and generated a maximum likelihood species tree in RAxML v.8.2.11 (Stamatakis, 2014) under GTR + Γ. We assessed the branch support with 500 rapid bootstrap (BS) replicates.

[1]www.anchoredphylogeny.com

Second, we generated a species tree under the multispecies coalescent using ASTRAL-III (Zhang et al., 2018). Using a batch Perl script, we generated individual gene trees using RAxML under GTR + Γ, assessing the branch support for each locus with 100 rapid BS replicates. We analyzed all the maximum likelihood gene trees in ASTRAL, measuring the branch support in two ways: by using the RAxML BS trees as input to ASTRAL with 100 replicates, and also by calculating the ASTRAL local posterior probability (LPP) (Sayyari and Mirarab, 2016) for each quadripartition.

To detail the gene tree conflict/support for each clade in the species tree, we used Phyparts (Smith et al., 2015). Phyparts takes an estimate of a species tree and set of rooted gene trees and provides four numbers for each clade: the number of loci supporting a clade, the number of loci supporting the main conflicting clade, the number of loci supporting all other conflicting clades, and the number of loci without information for a relationship. Trees were optimally rooted with our outgroups outside of Salviinae, but in cases where these were missing, we rooted trees with *Melissa* and/or *Lepechinia*. Gene trees that contained only *Salvia* were excluded from the Phyparts analysis. Note that rooting with Salviinae may inflate the gene tree support for the monophyly of *Salvia* and possibly also show misleading support or conflict for the relationships among *Lepechinia*, *Melissa*, and *Salvia*. However, we allowed this potentially incorrect rooting because our chief interest was in the relationships within *Salvia*. To mitigate the effects of uncertainty in the gene tree estimation providing artificial conflict (or support) for clades, we collapsed the branches in each gene tree with <33% BS.

Nuclear Dataset 1: Gene Tree Distances

We further examined the patterns of gene tree discordance to test if the observed gene tree discordance across *Salvia* and its constituent subgenera are consistent with the expectation under ILS alone. To do this, we first generated 1,000 gene trees under the multispecies coalescent using the *treesim.contained_coalescent* function in DendroPy v.4.5.2 (Sukumaran and Holder, 2010) using the ASTRAL species tree as the "true" tree. To compare the observed discordance with what would be expected under ILS, we measured the pairwise tree distance of each gene tree (expected and observed) from the ASTRAL species tree using three metrics, considering branching order alone and ignoring branch lengths: the Robinson–Foulds distance (RF) (Robinson and Foulds, 1981), the method proposed by Nye et al. (2006), and the clustering information (CI) metric proposed by Smith (2020a). Calculations were made on complete gene trees or gene trees pruned to the subgenus of interest, as appropriate. Since some of the observed gene trees were missing terminals, we only used observed gene trees which contained >75% of all terminals in the clade of interest. Gene tree distances were calculated using the "TreeDist" package in R (Smith, 2020b), collapsing all the branches in the observed gene trees with <33% BS. The distances were normalized so that they ranged from 0 to 1, with 0 indicating complete agreement between the gene tree and species tree. We tested for mean differences observed in the gene tree discordance among clades using a one-way ANOVA with *post hoc* testing using the Tukey Test with the *aov* and *glht* functions in the "stats" and "multcomp" (Hothorn et al., 2008) R packages, respectively. We tested for differences in the mean gene tree discordance between observed and expected gene trees using a two-tailed Welch's *t*-test.

Nuclear Dataset 2: Placeholder Dataset

Our second nuclear dataset investigated deeper phylogenetic relationships in *Salvia*, accounting for both ILS and horizontal gene flow. Because the existing methods for inferring phylogenetic networks are computationally demanding for datasets with more than several dozen terminals, we constructed a dataset of one representative for each subgenus. For each subgenus placeholder, we selected the sample with the greatest number of captured loci, and in the case of ties, the total number of aligned bp. We did not allow any missing data, yielding a matrix of 57 loci for 10 species of *Salvia* plus *Lepechinia chamaedryoides* (Balb.) Epling as the outgroup. To reconstruct the phylogenetic networks, we first generated concordance factors for each possible quartet. Using a batch script, we ran MrBayes v.3.2.6 (Ronquist et al., 2012) to find the best gene tree for each locus. The gene trees were inferred under GTR + I + Γ using three runs of three chains each for five million generations each with sampling every 5,000 generations with a chain temperature of 0.4, swap frequency of 500 generations, and a 30% burnin. Following the MrBayes analysis, a Bayesian concordance analysis on the posterior sample of gene trees was conducted in BUCKy v.1.4.4 (Ané et al., 2007; Larget et al., 2010) with 100,000 post-burnin generations and the amount of *a piori* discordance among loci set to the default of 1. This analysis calculates all possible quartets and prunes on the MrBayes gene trees to all but the four terminals of interest. Then, BUCKy is run on each pruned gene tree to generate a table of all quartet concordance factors (CFs) and their SEs. Using these CFs, we generated a preliminary population tree using Quartet MaxCut (Snir and Rao, 2012).

Using the BUCKy CFs and the Quartet MaxCut tree, we calculated a phylogenetic network with the SNaQ function in the Julia package PhyloNetworks (Solís-Lemus and Ané, 2016; Solís-Lemus et al., 2017). This package uses maximum pseudo-likelihood to fit a network while also accounting for ILS. PhyloNetworks considers quartet topologies only and does not take into account information from branch lengths in individual gene trees. Furthermore, PhyloNetworks assumes a level-1 network: a network where each hybrid node only has one lineage transferring genetic material horizontally. We first tested the fit of models allowing from 0–5 reticulation events (h) and compared the models using their pseudo-likelihood score. The best network model was selected by examining at which value of h the pseudo-likelihood score plateaus, following the recommendation of Solís-Lemus et al. (2017). For each value of h, we selected the best network over 30 search replicates. We examined the branch support on the best phylogenetic network using the bootsnaq function with 50 runs of 10 replicates each.

Plastome Assembly and Phylogenetic Analysis

We assembled the nearly complete plastomes of the Salviinae samples by mapping the off-target reads to previously published plastomes of *Salvia* for the ingroup or *Melissa* for the outgroup Salviinae. The assembly of the plastomes was conducted in Geneious v.10.2.3 (Kearse et al., 2012), following the procedure of Rose et al. (2021). For outgroup Salviinae, we used the whole plastome sequence of *Melissa yunnanensis* C.Y.Wu & Y.C.Huang (GenBank accession MT634148.1) as a reference. For *Salvia*, we constructed a "super" reference sequence based on the strict consensus of 18 GenBank plastomes (**Supplementary Data Sheet S2**) aligned with MAFFT v.7.023b (Katoh and Standley, 2013) under default parameters.

We used Geneious to map all the forward and reverse reads from our sequences by first trimming all raw reads, and then assembling them to the appropriate reference using an iterative refinement of up to five times with the default Geneious mapper and medium sensitivity. Consensus sequences were generated using the strict consensus approach. If the coverage for a particular site was <7, the consensus nucleotide was scored as a gap. Unmapped regions were treated as missing data and reads mapped to multiple positions were excluded from consensus calculations. Newly generated plastomes were aligned with the aforementioned GenBank sequences using MAFFT with default parameters. Ambiguously aligned regions were removed manually and were generally distinguished by putative inversions, repeat regions, an abundance of gaps, and/or uncertain base calls.

A plastome tree was inferred in RAxML under GTR + Γ with 500 rapid BS replicates. As described above in Section "Nuclear Dataset 1: Gene Tree Distances", we measured the tree-to-tree distances between the entire plastome tree and its subclades to the ASTRAL species tree.

RESULTS

Dataset Metrics

The aligned locus length for the 114 loci ranged from 105–3,671 bp, with a mean length of 1,133 bp. The samples contained sequence data for an average of 96.25 loci, with most locus dropout in the non-Salviinae outgroups. We were able to extract the majority of the plastome, with aligned plastomes totaling 157,683 bp.

Subgeneric Monophyly and Major Relationships in *Salvia*

We were able to root 101 of the 114 gene trees. Species trees resulting from concatenation and accounting for ILS with ASTRAL are completely congruent in the major backbone relationships in *Salvia*, although support for these relationships sometimes varies across the approach and support metrics (**Figure 1** and **Supplementary Figures S1–S4**). The ASTRAL normalized quartet score, or proportion of the gene tree quartet trees satisfied by the species tree, is 0.91, suggesting a clear underlying topology despite some discordance. The monophyly of *Salvia* is strongly supported by all measures of support (ASTRAL LPP = 1.0/ASTRAL BS = 100/concatenated BS = 100). In addition, the monophyly of each subgenus for which we had multiple samples is strongly supported by BS/LPP and by the vast majority of loci, with two exceptions. First, subg. *Heterosphace*, while unambiguously supported by measures of statistical support (ASTRAL LPP = 1.0/ASTRAL BS = 100/concatenated BS = 100), has 23/84 (27%) informative loci conflicting its monophyly. Second and more strikingly, the monophyly of subg. *Audibertia* is poorly supported by ASTRAL LPP (0.69) with more loci conflicting its monophyly than supporting it (55/73, 75%). However, its monophyly is more strongly supported by the other metrics (ASTRAL BS = 91/concatenated BS > 99), although both sections of subg. *Audibertia*: sects. *Audibertia* (*S. columbariae* Benth., *S. mellifera* Greene, *S. munzii* Epling) and *Echinosphace* (*S. californica* Brandegee, *S. funerea* M.E. Jones) are more strongly supported as monophyletic.

The earliest divergence in *Salvia* involves two major clades. First is a clade formed by the most recent common ancestor (MRCA) of subg. *Glutinaria* and *Calosphace* (ASTRAL LPP = 1.0/ASTRAL BS = 100/concatenated BS = 100). Subgenus *Glutinaria* is sister to all remaining subgenera, with a grade formed by the successive sisters of subg. *Zhumeria* and *Dorystaechas*, and subg. *Audibertia,* sister to subg. *Calosphace*. All of these relationships are strongly supported with the exception of the placement of subg. *Glutinaria* (ASTRAL LPP = 0.85/ASTRAL BS = 89/concatenated BS > 99; 22/71 informative gene trees).

The second major clade is formed by the MRCA of subg. *Perovskia* and *Salvia*. The monophyly of this clade is not fully supported (ASTRAL LPP = 0.73/ASTRAL BS = 92/concatenated BS = 100; 25/82 informative gene trees), nor are many of the intersubgeneric relationships within it (**Figure 1**). Subgenus *Perovskia* is sister to subg. *Rosmarinus* + *Heterosphace* + *Salvia* + *Sclarea* (ASTRAL LPP = 0.91/ASTRAL BS = 95/concatenated BS = 100; 22/69 informative gene trees). While the monophyly of subg. *Heterosphace* + *Salvia* + *Sclarea* is fully supported, relationships among the subgenera are slightly less certain, with subg. *Salvia* sister to *Sclarea* being the best resolution of relationships (ASTRAL LPP = 0.94/ASTRAL BS = 99/concatenated BS = 100; 26/81 informative gene trees).

The backbone of the plastome tree is nearly identical to that of the nuclear species trees (**Supplementary Figure S5**), with all major nodes and monophyly of the subgenera receiving maximal support except for the placement of subg. *Glutinaria* (BS = 89). The only major topological difference is that subg. *Perovskia* is weakly supported as sister to *Rosmarinus* (BS = 47).

Infrageneric Relationships, Shallow-Scale Resolution, and Gene Tree Discordance

Within subg. *Calosphace*, our nuclear data suggest that sect. *Axillares* is sister to the *Hastatae* clade, but with strongly conflicting support (ASTRAL LPP = 0.38/ASTRAL BS = 86/concatenated BS = 92; 20/60 informative gene trees),

while our plastid data place sect. *Axillares* as sister to all other *Calosphace* (BS = 100). There is also uncertainty about the deepest divergences in subg. *Salvia*, with weak support based on the ASTRAL and concatenated analyses.

Overall, there is fairly strong support (>90% support across all metrics) for many shallow-scale relationships, but support is notably very low or non-existent for some ASTRAL clades which do not appear in the best tree in the concatenated analysis or are in the low frequency in the BS replicates (**Supplementary Figure S6**), especially within the radiation of core *Calosphace* (e.g., relationships among *S. chamaedryoides* Cav., *S. coahuilensis* Fern., *S. microphylla* Kunth, and *S. muelleri* Epling), subg. *Salvia* (e.g., if *S. officinalis* s.s. is monophyletic or not), and subg. *Sclarea* (e.g., the placement of *S. sclarea* L.). There is a much more obvious infrageneric gene tree conflict between the nuclear loci and the plastome, with many shallower relationships conflicting between the two datasets, especially in subg. *Calosphace*, *Salvia*, and *Sclarea* (**Figure 2**).

Gene tree distances demonstrate significant gene tree conflicts within subgenera, both within and across genomic compartments (**Table 1**), and there are also significant differences in the gene tree distances among the subgenera for all three metrics (RF: $F_{5,454} = 145.7$, $p < 0.0001$; Nye: $F_{5,454} = 98.6$, $p < 0.0001$; CI: $F_{5,454} = 92.4$, $p < 0.0001$). *Post hoc* testing suggests that compared with the ASTRAL species tree, subg. *Glutinaria* and *Heterosphace* have significantly less discordant nuclear gene trees on average than all other subgenera, while subg. *Salvia* is more discordant than *Calosphace* for the Nye and CI metrics, and subg. *Sclarea* is more discordant than *Calosphace* for CI alone (**Table 1**). Likewise, although they are only point estimates, the discordance between the plastome and the ASTRAL tree is elevated for subg. *Sclarea* and especially, *Salvia* relative to all other subgenera (**Table 1**).

Compared with the expectation under ILS alone, the nuclear gene tree discordance in the observed gene trees is generally on par with or is less than what would be expected for under ILS based on RF distance but is greater than what would be expected in subg. *Calosphace*, *Salvia*, and *Sclarea* based on both the Nye and CI metrics (**Table 2**).

Phylogenetic Networks

The best phylogenetic network contained four reticulation events along the backbone of *Salvia* (hmax = 4). The major topology (i.e., bifurcating backbone) was identical to that recovered by the ASTRAL and concatenated analysis of all nuclear loci (**Figure 3**). Quartet CF along the backbone were generally > 0.50, and these edges received full BS support

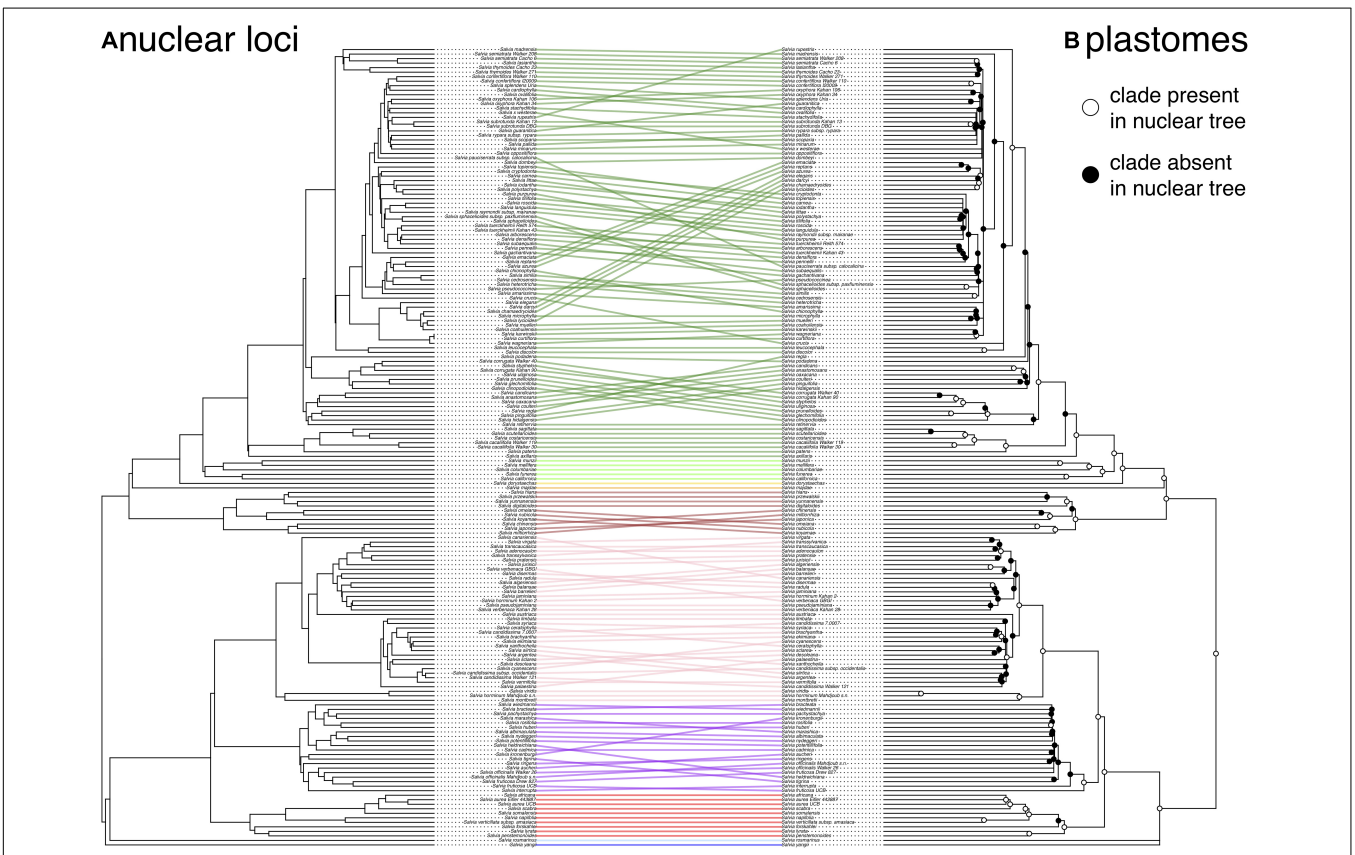


FIGURE 2 | Tanglegram illustrating the cytonuclear discordance in *Salvia* based on the nuclear ASTRAL species tree **(A)** and the plastome tree **(B)**. Links connect identical tips, with nodes rotated to minimize the link overlap. Ingroup links are colored by subgenus. Clades that differ between the two trees are indicated by filled circles on the plastome tree. To minimize the discordance caused by the error in the gene tree estimation, clades in the plastome tree with <75% bootstrap support have been collapsed into polytomies.

with the exception of the placement of subg. *Zhumeria* (CF = 0.42, BS = 96) and the sister relationship of subg. *Salvia* and *Sclarea* (CF = 0.42, BS = 86). While the BS analysis found evidence for horizontal gene flow, there was considerable uncertainty regarding the number and placement of reticulation events, with BS replicates recovering either three (52%) or four reticulation events. Inheritance probabilities (γ, the fraction of the nuclear genome involved in a reticulation event) for the four reticulation events on the best-fitting network ranged from 7 to 36% (**Figure 3**). The best network recovered gene flow from the stem of subg. *Salvia* to *Heterosphace* (γ = 0.36, BS = 66), stem subg. *Rosmarinus* to stem MRCA of *Calosphace* + *Glutinaria* (γ = 0.26, BS = 66), stem MRCA of subg. *Calosphace* + *Glutinaria* to *Zhumeria* (γ = 0.10, BS = 10), and stem subg. *Dorystaechas* to *Audibertia* (γ = 0.07, BS = 34). Alternative reticulation events found in frequency > 10% in the BS replicates involved the stem of subg. *Salvia* and *Sclarea* (BS = 22, with alternative relationships among *Heterosphace*, *Salvia*, and *Sclarea*), stem MRCA subg. *Rosmarinus* + *Salvia* and MRCA *Calosphace* + *Glutinaria* (BS = 28), stem subg. *Glutinaria* and stem *Audibertia* + *Calosphace*

TABLE 1 | Summary of the mean gene tree distances in *Salvia* and the selected subgenera within and across genomes.

clade	*n*-tips	nuclear loci			plastome		
		RF	Nye	CI	RF	Nye	CI
Salvia-wide	179	0.61[b]	0.41[b]	0.36[b]	0.63	0.37	0.29
subg. *Calosphace*	93	0.69[c]	0.52[d]	0.57[d]	0.72	0.41	0.45
subg. *Glutinaria*	10	0.29[a]	0.20[a]	0.24[a]	0.38	0.23	0.25
subg. *Heterosphace*	10	0.28[a]	0.18[a]	0.22[a]	0.08	0.08	0.07
subg. *Salvia*	37	0.67[bc]	0.44[bc]	0.50[c]	0.93	0.64	0.74
subg. *Sclarea*	20	0.70[c]	0.48[cd]	0.49[c]	0.78	0.46	0.51

The three different distance metrics, Robinson–Foulds (RF), Nye, and clustering information (CI), compare the topology of the gene trees to the ASTRAL species tree, and we considered the branching order alone. Metrics were calculated on the gene trees that can be rooted and have clade occupancy > 75% of sampled tips. The distances were normalized so that they ranged from 0 to 1, with 0 indicating complete agreement between a gene tree and a species tree. The letters denote significantly different among-group differences for each metric in mean nuclear gene tree discordance compared to the species tree based on an ANOVA with post hoc testing at α = 0.05. Differences among the plastomes were not tested because they represent one gene tree.

TABLE 2 | Nuclear gene tree distances in *Salvia* and selected subgenera.

Clade	RF	Nye	CI	t_{RF}	p_{RF}	t_{NYE}	p_{NYE}	t_{CI}	p_{CI}
	mean (SD)	mean (SD)	Mean (SD)						
***Salvia*-wide**									
Observed	0.61 (0.07)	0.41 (0.06)	0.36 (0.06)	**–4.87**	**7.82×10^{-6}**	**7.84**	**7.45×10^{-11}**	**5.08**	**3.65×10^{-6}**
Expected	0.65 (0.03)	0.35 (0.02)	0.32 (0.02)						
subg. *Calosphace*									
Observed	0.69 (0.08)	0.52 (0.10)	0.57 (0.11)	**–2.92**	**4.39×10^{-3}**	**10.67**	**$< 1.00 \times 10^{-15}$**	**8.83**	**1.12×10^{-13}**
Expected	0.72 (0.05)	0.40 (0.03)	0.46 (0.04)						
subg. *Glutinaria*									
Observed	0.29 (0.21)	0.20 (0.17)	0.24 (0.18)	–1.04	0.30	0.49	0.63	–0.029	0.98
Expected	0.32 (0.17)	0.19 (0.09)	0.24 (0.11)						
subg. *Heterosphace*									
Observed	0.28 (0.17)	0.18 (0.11)	0.22 (0.13)	–1.06	0.29	0.98	0.33	–0.11	0.91
Expected	0.30 (0.17)	0.17 (0.09)	0.22 (0.11)						
subg. *Salvia*									
Observed	0.67 (0.17)	0.44 (0.16)	0.50 (0.16)	–0.91	0.36	**4.70**	**8.84×10^{-6}**	**2.56**	**0.012**
Expected	0.69 (0.12)	0.36 (0.07)	0.46 (0.09)						
subg. *Sclarea*									
Observed	0.70 (0.10)	0.48 (0.10)	0.49 (0.11)	**–3.10**	**2.62×10^{-3}**	**8.35**	**1.14×10^{-12}**	**4.37**	**3.51×10^{-5}**
Expected	0.74 (0.07)	0.38 (0.04)	0.43 (0.05)						

The three different distance metrics, RF, Nye, and CI, compare the topology of the gene trees to the ASTRAL species tree, and we considered the branching order alone. The gene trees are either empirical trees that can be rooted and have clade occupancy > 75% of sampled tips (observed) or 1,000 gene trees simulated under the multispecies coalescent using the ASTRAL species tree (expected). The distances were normalized so that they ranged from 0 to 1, with 0 indicating complete agreement between a gene tree and the species tree. The t-values and associated p-values for each distance metric/clade combination are based on Welch's t-test with the hypothesis that the mean tree distance for the observed and expected gene trees are equal, or in other words, that the tree distances based on empirical data are what would be expected under incomplete lineage sorting alone. Significant t/p-values at α = 0.05 are indicated in bold.

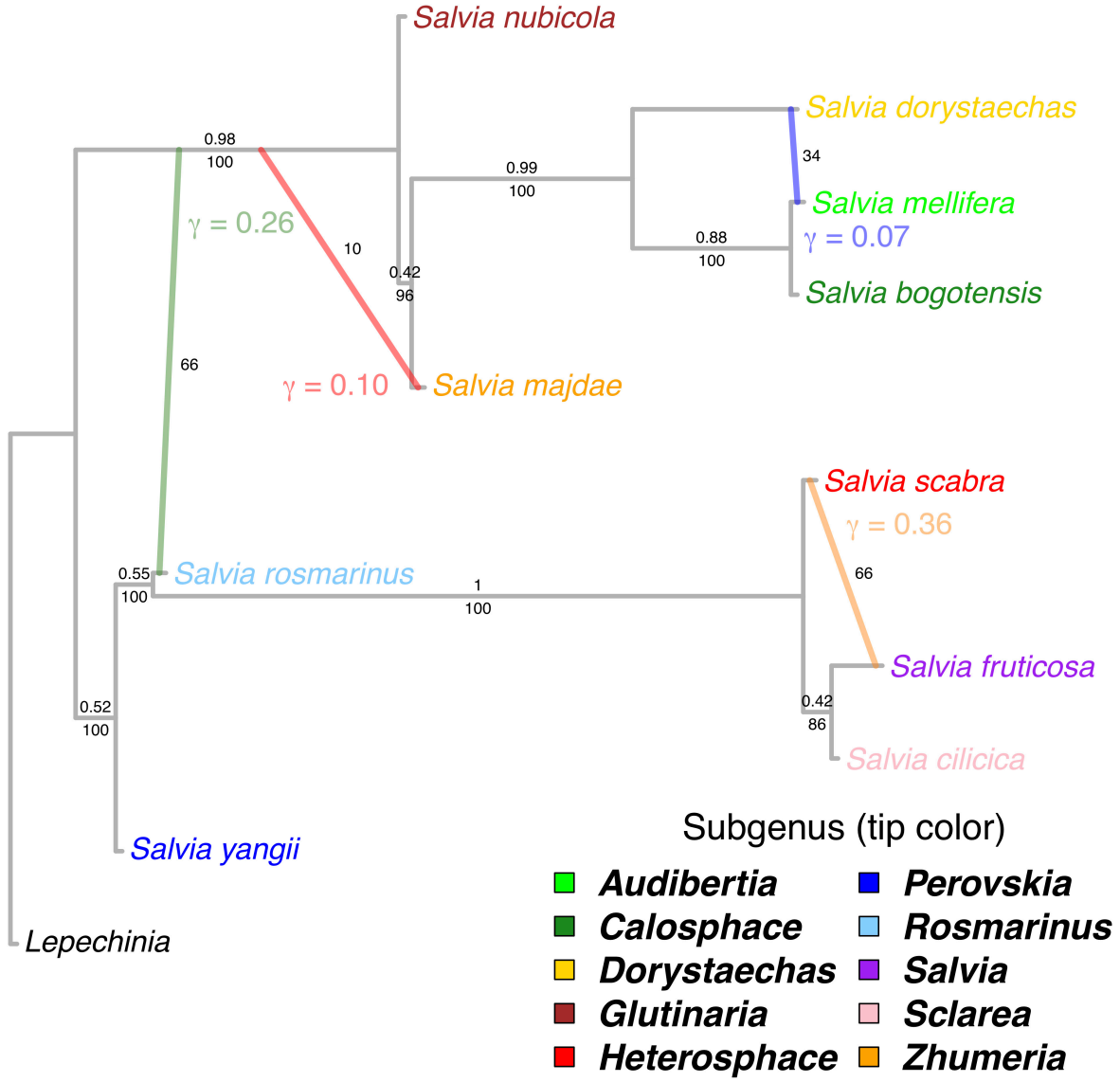


FIGURE 3 | Phylogenetic network based on 57 nuclear genes depicting backbone relationships in *Salvia* with one exemplar per subgenus, with internal branch lengths in coalescent units. The best-fitting network has four reticulation events. Inheritance probabilities (γ) are indicated next to the lineage which is inferred to have received genetic material. The numbers above the branches on the bifurcating major topology are quartet concordance factors or the proportion of the genome supporting each quartet. The numbers below the major topology branches and to the right of hybrid branches are bootstrap support values. Ingroup tips are colored by subgenus.

(BS = 40), and stem subg. *Glutinaria* and stem MRCA *Calosphace* + *Dorystaechas* (BS = 40).

DISCUSSION

A Robust Phylogenetic Hypothesis for *Salvia*

Our results demonstrate that despite the gene tree discordance, the backbone relationships of *Salvia* are identical using nuclear and plastid data. These genomes support the monophyly of all currently recognized subgenera and are largely concordant regarding the intersubgeneric relationships, where supported. Discordance among nuclear loci can largely be reconciled by invoking ILS and horizontal gene flow.

Our results largely corroborate our previous analysis involving a larger number of loci with fewer terminals (Kriebel et al., 2019). Despite the fewer loci examined in this study, the average locus length is nearly twice as long (623 vs. 1,133 bp), presumably increasing the accuracy in the gene tree estimation. Additionally, compared with Kriebel et al. (2019) we found increased ASTRAL BS support for the monophyly of subg. *Audibertia* (>0.99 vs. 0.79) and for the sister relationship of subg. *Salvia* and *Sclarea* (0.99 vs. 0.79). However, we did not find convincing support for

subg. *Audibertia* based on the ASTRAL LPP, and the monophyly for this subgenus based on the nuclear data has been somewhat unclear (Walker et al., 2004, 2015; Walker and Sytsma, 2007; Drew et al., 2017; Will and Claßen-Bockhoff, 2017), although it is clearly monophyletic based on the datasets with a good sampling of plastid loci (Walker et al., 2015; **Supplementary Figure S5**).

Unexpectedly, we found some uncertainty regarding the placements of subg. *Perovskia*, *Rosmarinus*, and *Zhumeria*, which were previously placed with BS = 100 in Kriebel et al. (2019). Nevertheless, the placement of all these subgenera, especially *Perovskia* and *Rosmarinus*, has varied widely across previous molecular studies incorporating low-copy nuclear loci (Drew et al., 2017), transcriptomes (Mint Evolutionary Genomics Consortium, 2018), nuclear ribosomal ITS/ETS (Drew and Sytsma, 2012; Will and Claßen-Bockhoff, 2017; Kriebel et al., 2019), and plastid data (Walker and Sytsma, 2007; Drew and Sytsma, 2012; Drew et al., 2017; Will and Claßen-Bockhoff, 2017; Zhao et al., 2020, 2021). The varying placement of subg. *Rosmarinus* across studies–and indeed across the loci examined in this study–can be explained by a combination of ILS and ancient horizontal gene flow (see below). This may be true for subg. *Zhumeria* as well, although it is less likely that horizontal gene flow has been involved in that case given the low BS support for any such gene flow.

While the backbone topology of the plastome of *Salvia* does not contradict that of the nuclear tree, it is still incompletely supported. However, it is unclear how much more information about the major relationships in *Salvia*, especially the relationships of subg. *Perovskia* and *Rosmarinus*, can be garnered from it. Analyses incorporating complete or nearly complete plastomes have failed to recover a fully supported backbone (Zhao et al., 2020; **Supplementary Figure S5**). Therefore, adding the portions of the plastome that we excluded does not seem to provide a viable solution. Mitogenomes, possibly combined with plastomes, are a possible avenue of research for a fully supported organellar phylogeny.

A final open question regarding the deeper-level phylogeny of *Salvia* is the relationship of the unsampled subg. *Meriandra*. We expect this subgenus to be closely related to subg. *Dorystaechas* and possibly even sister to it based on previous molecular results (Walker and Sytsma, 2007; Drew and Sytsma, 2012; Will and Claßen-Bockhoff, 2017; Kriebel et al., 2019). The placement of subg. *Meriandra* has important implications, not only for the historical biogeography of the genus but also implications for the timing and geographic location of any gene flow between the ancestors of subg. *Audibertia* and *Dorystaechas*, if present (see below).

Evidence for Gene Flow in the Backbone of *Salvia*: But Where?

While the major topology of our phylogenetic network in *Salvia* is clear and strongly recovers the same bifurcating backbone found across other analyses, each with different assumptions, it also suggests that such a relatively simple tree may not be the best model of the phylogenetic history of *Salvia* (**Figure 3**). While all of our BS trees recovered at least three gene flow events, there is considerable uncertainty regarding which clades were involved in some of the horizontal gene flow events. We are, however, fairly certain that one gene flow event involved the stem MRCA of subg. *Glutinaria* and *Calosphace*, with the gene flow involving either the ancestor of subg. *Rosmarinus* alone (BS = 66) or the MRCA of subg. *Rosmarinus* and *Salvia* (BS = 28), and this likely explains the uncertainty regarding the placement of subg. *Perovskia* and *Rosmarinus*.

Likewise, it seems probable that the uncertainty regarding the branching order of subg. *Heterosphace*, *Salvia*, and *Sclarea* is the result of the gene flow. Although our phylogenetic network favors horizontal gene flow between subg. *Heterosphace* and *Salvia* as a better explanation for the discordance (BS = 66), an alternative resolution of the relationships with the gene flow between subg. *Salvia* and *Sclarea* is also possible (BS = 22). Given this result, it is also possible that the constraint of the level-1 network is not an appropriate model, and subg. *Salvia* may be a hybrid between subg. *Heterosphace* and *Sclarea*. This hypothesis requires further testing.

On the other hand, we found poor support for the remaining two inferred horizontal gene flow events on our best network, especially for the gene flow involving subg. *Zhumeria*. While slightly better supported, the horizontal gene flow between subg. *Audibertia* and *Dorystaechas* seem implausible since it necessitates the gene flow between their ancestors in North America and Southwest Asia (Kriebel et al., 2019). Despite its absence from the best network, the BS replicates suggest a strong possibility of gene flow involving subg. *Glutinaria*, especially with the MRCA of *Audibertia* + *Calosphace* (possibly extended to *Dorystaechas*), which would be more consistent with our current understanding of the historical biogeography of *Salvia*.

Apart from a level-1 network possibly being an unreasonable restriction to our dataset, the uncertainty in the placement of horizontal gene flow may be due to the relatively few loci employed here. For example, SNAq may recover false positive hybridization events in the datasets with < 100 loci (Solís-Lemus and Ané, 2016). Overall, our results highlight that it is essential to complement searches for best-fitting networks with BS analyses.

Support for Key Infrageneric Structure

Our AHE data provides good support for many shallow-scale relationships in *Salvia* (**Figure 1** and **Supplementary Figures S1–S3**). We clarify the placement of sect. *Axillares* within subg. *Calosphace*, with the nuclear evidence slightly favoring a hypothesis of the sister relationship of sect. *Axillares* and the *Hastatae* clade (sects. *Blakea*, *Hastatae*, and *Standleyana*), although the support based on ASTRAL LPP is noticeably weak. This relationship is identical to that suggested by the nuclear ribosomal DNA (Fragoso-Martínez et al., 2018; Kriebel et al., 2019, Appendix S7), but not the plastid data (Drew et al., 2017; Will and Claßen-Bockhoff, 2017; Fragoso-Martínez et al., 2018) or one low copy nuclear marker (*PPR*: Drew et al., 2017), which instead placed sect. *Axillares* as sister to the remainder of subg. *Calosphace*. It is still unclear if the uncertainty in the placement is due to ILS or gene flow, which should be investigated in future studies focused on subg. *Calosphace*,

especially since the placement of sect. *Axillares* has important macroevolutionary implications.

Where our sampling of the Old World lineages permits, we corroborated relationships among deep subgeneric splits which were strongly supported in previous studies (subg. *Glutinaria*: Hu et al., 2018, subg. *Heterosphace*: Will and Claßen-Bockhoff, 2014, 2017). Within the other Old World subgenera, the support for relationships in previous studies has not been robust enough to warrant a discussion of the major relationships (Will and Claßen-Bockhoff, 2014, 2017), and thus our results, where it is well supported, are novel.

From a taxonomic perspective, it is encouraging that for the few species for which we have multiple accessions, morphologically defined species are monophyletic, paraphyletic due to the inclusion of only one other species, or essentially form a polytomy with other morphologically similar species, rather than polyphyletic and/or found in large polytomies (**Figure 1** and **Supplementary Figures S1–S3**). This suggests that our AHE dataset has the power to not only resolve deeper relationships in *Salvia* but also to provide information pertinent to species delimitation.

Infrageneric Structure: Incomplete Lineage Sorting and Horizontal Gene Flow Explain Strong Gene Tree Discordance

While phylogenomic datasets show great promise to resolve relationships in previously intractable angiosperm lineages, the irony is that many of these groups have undergone rapid radiations (Larson et al., 2020; Shee et al., 2020; Rose et al., 2021; Thomas et al., 2021), which increases the chance of gene tree heterogeneity due to ILS (Pamilo and Nei, 1988; Maddison, 1997; Oliver, 2013). Rampant gene tree discordance need not mean that species trees are poorly supported, provided that the discordance is consistent with the underlying model used to generate the species tree. Indeed, our analysis suggests that much of the gene tree discordance is at least consistent with ILS, given the high support for many infrageneric relationships based on our ASTRAL analysis (**Figure 1** and **Supplementary Figures S1, S2**). Conversely, the low support for relationships in approaches that only take ILS into account may be due to either the lack of information about a given relationship in the underlying sequence data, or a more complex model of relationships (i.e., one involving horizontal gene flow). Our results demonstrate that while the average discordance of nuclear gene trees is consistent with what would be expected under ILS alone in the relatively under-sampled subg. *Glutinaria* and *Heterosphace*, it exceeds what would be expected under ILS in subg. *Calosphace*, *Salvia*, and *Sclarea* (**Table 2**). More strikingly, since that under ILS gene tree discordance should increase simply as a function of taxon sampling, it is notable that subg. *Salvia* and *Sclarea* have observed mean nuclear gene tree discordance on par with or slightly lower than that for subg. *Calosphace*, despite being represented by many fewer tips (**Table 1**).

One possible explanation for this is that the increased ILS results from the very rapid radiations of these clades, in combination with much younger crown ages for the MRCAs of what this study samples in subg. *Sclarea* (13.4 My) and *Salvia* ($\leq$7.8 My) relative to *Calosphace* (20.1 My) (Kriebel et al., 2019). However, based on the excess of the nuclear gene tree discordance in the aforementioned clades relative to the expectation under the multispecies coalescent, we suggest that in these clades, especially in subg. *Salvia* and *Sclarea*, the multispecies coalescent does not provide an ideal model of phylogenetic relationships. Instead, a model with horizontal gene flow in these lineages is likely a better explanation for the excess of gene tree discordance observed in our data. While another possibility for this pattern is that error in the gene tree estimation adds artificial discordance, we reject this as a major complicating factor given that we collapsed very poorly supported edges in observed gene trees.

The stark discordance present between the ASTRAL tree and the plastome tree at many shallow nodes in subg. *Calosphace*, *Salvia*, and *Sclarea*, with especially large distances between the nuclear and plastid trees in subg. *Salvia,* is also highly suggestive of an important contribution from horizontal gene flow, either hybridization or introgression (**Figure 2**). However, depending on the amount of past backcrossing, a signal for past gene flow may be absent from the nuclear genome in some cases. While we did not test it here, we do not think ILS is a likely explanation for the intergenomic gene tree conflict given the results from other angiosperm systems (Folk et al., 2017; Morales-Briones et al., 2018; Lee-Yaw et al., 2019; Rose et al., 2021), although error in the species tree estimation is a possible explanation. In future studies, we expect that relatively under-sampled subgenera should show increasing levels of cytonuclear discordance as we increase species sampling, especially within the subg. *Glutinaria* (Hu et al., 2018). The potentially confounding effects of polyploidy and whole-genome duplications (WGD) were not evaluated here but are being investigated.

Finally, it is worth a brief note concerning why we found that the mean RF tree distance often conflicts with the other distance metrics by demonstrating that the mean discordance is either on par with expectations under the multispecies coalescent or observed gene trees are, in some cases, less discordant. Despite being a widely used metric, RF distance is probably too conservative in penalizing against relatively minor topological differences (Smith, 2020a), as the movement of a single tip may result in maximum tree-to-tree distances even though all other tips show the same branching pattern. Thus, collapsing poorly supported edges of observed gene trees into polytomies downplays discordance, while any minor topological differences in the fully-resolved expected gene trees are penalized.

CONCLUSION

Our updated AHE dataset provides evidence for a well-supported backbone of *Salvia* and indicates that there is an emerging consensus of relationships in the genus that extends across genomic compartments. Past difficulty in inferring relationships has likely been caused by a combination of uninformative markers, ILS, and horizontal gene flow. To

the latter point, while our dataset clearly shows evidence of horizontal gene flow at deep and shallow scales in *Salvia*, we are presently unable to confidently demonstrate how many ancient gene flow events occurred and where they are placed. This highlights the importance of assessing the support for the best-fitting phylogenetic network, rather than only presenting the best network.

Several issues still need clarification, especially in the placement of subg. *Meriandra* and in the monophyly of subg. *Audibertia*. We are confident that future analyses using this same or expanded set of loci, in concert with an evaluation of polyploidy and WGD processes, will resolve these issues. Additionally, targeted analyses of clades or further methodological advances will allow us to tease apart horizontal gene flow at shallower scales. Our phylogenetic hypothesis, as well as future, time-calibrated phylogenetic hypotheses of the entire genus *Salvia*, its constituent subgenera, and targeted clades, will provide an invaluable framework for which to conduct multiple comparative analyses in this fascinating genus.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the NCBI Sequence Read Archive (SRA) as BioProject PRJNA773953.

AUTHOR CONTRIBUTIONS

JR, RK, BD, and KS conceived and undertook the project. JR, RK, LK, AD, JG-G, FC, EL, AL, and BD assisted with the data collection. JR analyzed the data and led the writing with contributions from all authors. All the authors contributed to the article and approved the submitted version.

FUNDING

This manuscript was funded in part by the University of Wisconsin Botany Department Hofmeister Endowment, NSF-DEB collaborative grant to KS and BD (DEB-1655606 and DEB-1655611), and TUBITAK project number 2219 to FC for postdoctoral studies in the United States. JG-G appreciates the financial support provided by CONACYT by means of the project CB-2015-01-255165.

ACKNOWLEDGMENTS

We gratefully acknowledge Holly Forbes from the UC-Berkeley Botanical Garden, and Cindy Newlander and Mike Kintgen from the Denver Botanical Garden for granting permission to collect garden specimens and assisting us. We are grateful to Jay Walker, N. Ivalú Cacho, Eleftherios Dariotis, and Rolando Uría for their help in collecting the specimen.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.767478/full#supplementary-material

Supplementary Figure S1 | The ASTRAL species tree of *Salvia* and outgroups, with local posterior probabilities on branches. Ingroup branches are colored by subgenus.

Supplementary Figure S2 | The ASTRAL species tree of *Salvia* and outgroups, with bootstrap support on branches. Ingroup branches are colored by subgenus.

Supplementary Figure S3 | The RAxML maximum likelihood tree of *Salvia* and outgroups based on the concatenated nuclear matrix, with bootstrap support on branches. Ingroup branches are colored by subgenus.

Supplementary Figure S4 | Phyparts summary of gene trees. Pies at major nodes summarize the percentage of various phylogenetic signals across 101 gene trees which can be rooted. The numbers at the left of the pies show the total number of gene trees in which the clade is found, followed by the total number of gene trees that conflict with that clade. The remainder of the gene trees, if any, do not provide information on that particular relationship. Ingroup branches are colored by subgenus.

Supplementary Figure S5 | The RAxML maximum likelihood tree of *Salvia* and outgroups based on entire plastomes, with bootstrap support on branches. GenBank accessions are removed so that the plastome tree matches the nuclear tree in the tip composition. Ingroup branches are colored by subgenus.

Supplementary Figure S6 | Tanglegram illustrating the disagreement between the ASTRAL **(A)** and concatenated maximum likelihood **(B)** species trees based on nuclear data. Links connect identical tips, with nodes rotated to minimize link overlap. Ingroup branches are colored by subgenus. Clades that differ between the two trees are indicated by filled circles on the concatenated maximum likelihood tree.

REFERENCES

Ané, C., Larget, B., Baum, D. A., Smith, S. D., and Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* 24, 412–426. doi: 10.1093/molbev/msl170

Baum, D. A. (2007). Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon* 56, 417–426. doi: 10.1002/tax.562013

Buddenhagen, C. E., Lemmon, A. R., Lemmon, E. M., Bruhl, J., Cappa, J., Clement, W. L., et al. (2016). Anchored phylogenomics of angiosperms I: assessing the robustness of phylogenetic estimates. *BioRxiv* 086298. doi: 10.1101/086298

Celep, F., Atalay, Z., Dikmen, F., Doğan, M., Sytsma, K. J., and Claßen-Bockhoff, R. (2020). Pollination ecology, specialization, and genetic isolation in sympatric bee-pollinated Salvia (Lamiaceae). *Intl. J. Plant Sci.* 181, 800–811. doi: 10.1086/710238

Claßen-Bockhoff, R., Speck, T., Tweraser, E., Wester, P., Thimm, S., and Reith, M. (2004). The staminal lever mechanism in Salvia L. (Lamiaceae): a key innovation for adaptive radiation? *Org. Diver. Evol.* 4, 189–205. doi: 10.1016/j.ode.2004.01.004

Claßen-Bockhoff, R., Wester, P., and Tweraser, E. (2003). The staminal lever mechanism in Salvia L. (Lamiaceae) - a review. *Plant Biol.* 5, 33–41. doi: 10.1093/aob/mcr011

Degnan, J. H. (2018). Modeling hybridization under the network multispecies coalescent. *Syst. Biol.* 67, 786–799. doi: 10.1093/sysbio/syy040

Dizkirici, A., Celep, F., Kansu, C., Kahraman, A., Dogan, M., and Kaya, Z. (2015). A molecular phylogeny of Salvia euphratica sensu lato (Salvia L., Lamiaceae) and

its closely related species with a focus on the section Hymenosphace. *Plant Syst. Evol.* 301, 2313–2323. doi: 10.1007/s00606-015-1230-1

Drew, B. T., González-Gallegos, J. G., Xiang, C. L., Kriebel, R., Drummond, C. P., Walker, J. B., et al. (2017). Salvia united: the greatest good for the greatest number. *Taxon* 66, 133–145. doi: 10.12705/661.7

Drew, B. T., and Sytsma, K. J. (2011). Testing the monophyly and placement of Lepechinia in the tribe Mentheae (Lamiaceae). *Syst. Bot.* 36, 1038–1049.

Drew, B. T., and Sytsma, K. J. (2012). Phylogenetics, biogeography, and staminal evolution in the tribe Mentheae (Lamiaceae). *Am. J. Bot.* 99, 933–953. doi: 10.3732/ajb.1100549

Drew, B. T., and Sytsma, K. J. (2013). The South American radiation of Lepechinia (Lamiaceae): phylogenetics, divergence times and evolution of dioecy. *Bot. J. Linn. Soc.* 171, 171–190. doi: 10.1111/j.1095-8339.2012.01325.x

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., et al. (2016). Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Mol. Phylogen. Evol.* 94, 447–462. doi: 10.1016/j.ympev.2015.10.027

Folk, R. A., Mandel, J. R., and Freudenstein, J. V. (2017). Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Syst. Biol.* 66, 320–337. doi: 10.1093/sysbio/syw083

Folk, R. A., Soltis, P. S., Soltis, D. E., and Guralnick, R. (2018). New prospects in the detection and comparative analysis of hybridization in the tree of life. *Am. J. Bot.* 105, 364–375. doi: 10.1002/ajb2.1018

Fragoso-Martínez, I., Martínez-Gordillo, M., Salazar, G. A., Sazatornil, F., Jenks, A. A., García-Peña, M. D. R., et al. (2018). Phylogeny of the Neotropical sages (Salvia subg. Calosphace; Lamiaceae) and insights into pollinator and area shifts. *Plant Syst. Evol.* 304, 43–55.

Fragoso-Martínez, I., Salazar, G. A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Lemmon, E. M., et al. (2017). A pilot study applying the plant Anchored Hybrid Enrichment method to New World sages (Salvia subgenus Calosphace. Lamiaceae). *Mol. Phylogen. Evol.* 117, 124–134. doi: 10.1016/j.ympev.2017.02.006

González-Gallegos, J. G., Bedolla-García, B. Y., Cornejo-Tenorio, G., Fernández-Alonso, J. L., Fragoso-Martínez, I., García-Peña, M. D. R., et al. (2020). Richness and distribution of Salvia subg. Calosphace (Lamiaceae). *Intl. J. Plant Sci.* 181, 831–856. doi: 10.1086/709133

Hejase, H. A., and Liu, K. J. (2016). A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinform.* 17:422. doi: 10.1186/s12859-016-1277-1

Heled, J., and Drummond, A. J. (2009). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27, 570–580. doi: 10.1093/molbev/msp274

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biom. J.* 50, 346–363. doi: 10.1002/bimj.200810425

Hu, G. X., Takano, A., Drew, B. T., Liu, E. D., Soltis, D. E., Soltis, P. S., et al. (2018). Phylogeny and staminal evolution of Salvia (Lamiaceae, Nepetoideae) in East Asia. *Annals Bot.* 122, 649–668. doi: 10.1093/aob/mcy104

Jenks, A. A., Walker, J. B., and Kim, S. C. (2013). Phylogeny of new world Salvia subgenus Calosphace (Lamiaceae) based on cpDNA (psb A-trn H) and nrDNA (ITS) sequence data. *J. Plant Res.* 126, 483–496. doi: 10.1007/s10265-012-0543-1

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kriebel, R., Drew, B. T., Gonzáález-Gallegos, J. G., Celep, F., Antar, G. M., Pastore, J. F. B., et al. (2021). Stigma shape shifting in sages (*Salvia*: Lamiaceae) – hummingbirds guided the evolution of New World floral features. *Bot. J. Linn. Soc.* (in press).

Kriebel, R., Drew, B., González-Gallegos, J. G., Celep, F., Heeg, L., Mahdjoub, M. M., et al. (2020). Pollinator shifts, contingent evolution, and evolutionary constraint drive floral disparity in Salvia (Lamiaceae): evidence from morphometrics and phylogenetic comparative methods. *Evolution* 74, 1335–1355. doi: 10.1111/evo.14030

Kriebel, R., Drew, B. T., Drummond, C. P., González-Gallegos, J. G., Celep, F., Mahdjoub, M. M., et al. (2019). Tracking temporal shifts in area, biomes, and pollinators in the radiation of Salvia (sages) across continents: leveraging anchored hybrid enrichment and targeted sequence data. *Am. J. Bot.* 106, 573–597. doi: 10.1002/ajb2.1268

Larget, B. R., Kotha, S. K., Dewey, C. N., and Ané, C. (2010). BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* 26, 2910–2911. doi: 10.1093/bioinformatics/btq539

Larson, D. A., Walker, J. F., Vargas, O. M., and Smith, S. A. (2020). A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. *Am. J. Bot.* 107, 773–789. doi: 10.1002/ajb2.1469

Leaché, A. D., Harris, R. B., Rannala, B., and Yang, Z. (2014). The influence of gene flow on species tree estimation: a simulation study. *Syst. Biol.* 63, 17–30. doi: 10.1093/sysbio/syt049

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., and Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (Helianthus). *N. Phytol.* 221, 515–526. doi: 10.1111/nph.15386

Lemmon, A. R., and Lemmon, E. M. (2012). High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Syst. Biol.* 61, 745–761. doi: 10.1093/sysbio/sys051

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.* 46, 523–536.

Mint Evolutionary Genomics Consortium (2018). Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae. *Mol. Plant* 11, 1084–1096. doi: 10.1016/j.molp.2018.06.002

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30, i541–i548.

Mitchell, N., Lewis, P. O., Moriarty Lemmon, E., Lemmon, A. R., and Holsinger, K. E. (2017). Anchored phylogenomics resolves the evolutionary relationships in the rapid radiation of Protea L. (Proteaceae). *Am. J. Bot.* 104, 102–115. doi: 10.3732/ajb.1600227

Morales-Briones, D. F., Liston, A., and Tank, D. C. (2018). Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus Lachemilla (Rosaceae). *N. Phytol.* 218, 1668–1684. doi: 10.1111/nph.15099

Nye, T. M., Lio, P., and Gilks, W. R. (2006). A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics* 22, 117–119. doi: 10.1093/bioinformatics/bti720

Oliver, J. C. (2013). Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* 67, 1823–1830. doi: 10.1111/evo.12047

Pamilo, P., and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583.

Rieseberg, L. H., and Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trends Plants* 5, 65–84.

Robinson, D. F., and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Rose, J. P., Toledo, C. A., Lemmon, E. M., Lemmon, A. R., and Sytsma, K. J. (2021). Out of sight, out of mind: widespread nuclear and plastid-nuclear discordance in the flowering plant genus Polemonium (Polemoniaceae) suggests widespread historical gene flow despite limited nuclear signal. *Syst. Biol.* 70, 162–180. doi: 10.1093/sysbio/syaa049

Sayyari, E., and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* 33, 1654–1668. doi: 10.1093/molbev/msw079

Shee, Z. Q., Frodin, D. G., Cámara-Leret, R., and Pokorny, L. (2020). Reconstructing the complex evolutionary history of the Papuasian Schefflera radiation through herbariomics. *Front. Plant Sci.* 11:258. doi: 10.3389/fpls.2020.00258

Smith, M. R. (2020a). Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics* 36, 5007–5013. doi: 10.1093/bioinformatics/btaa614

Smith, M. R. (2020b). TreeDist: distances Between Phylogenetic Trees. R package version 2.0.3. Comprehensive R Archive Network. doi: 10.5281/zenodo.3528124

Smith, R. L., and Sytsma, K. J. (1990). Evolution of Populus nigra (sect. Aigeiros): introgressive hybridization and the chloroplast contribution of Populus alba (sect. Populus). *Am. J. Bot.* 77, 1176–1187.

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* 15:150. doi: 10.1186/s12862-015-0423-0

Snir, S., and Rao, S. (2012). Quartet MaxCut: a fast algorithm for amalgamating quartet trees. *Mol. Phylogenet. Evol.* 62, 1–8.

Solís-Lemus, C., and Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 12:e1005896. doi: 10.1371/journal.pgen.1005896

Solís-Lemus, C., Bastide, P., and Ané, C. (2017). PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.* 34, 3292–3298.

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

Sukumaran, J., and Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 26, 1569–1571.

Thomas, A. E., Igea, J., Meudt, H. M., Albach, D. C., Lee, W. G., and Tanentzap, A. J. (2021). Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand Veronica. *Am. J. Bot.* 108, 1289–1306.

Walker, J. B., Drew, B. T., and Sytsma, K. J. (2015). Unravelling species relationships and diversification within the iconic California Floristic Province sages (Salvia subgenus Audibertia, Lamiaceae). *Syst. Bot.* 40, 826–844.

Walker, J. B., and Sytsma, K. J. (2007). Staminal evolution in the genus Salvia (Lamiaceae): molecular phylogenetic evidence for multiple origins of the staminal lever. *Annals Bot.* 100, 375–391.

Walker, J. B., Sytsma, K. J., Treutlein, J., and Wink, M. (2004). Salvia (Lamiaceae) is not monophyletic: implications for the systematics, radiation, and ecological specializations of Salvia and tribe Mentheae. *Am. J. Bot.* 91, 1115–1125.

Wester, P., and Claßen-Bockhoff, R. (2007). Floral diversity and pollen transfer mechanisms in bird-pollinated Salvia species. *Ann. Bot.* 100, 401–421.

Will, M., and Claßen-Bockhoff, R. (2014). Why Africa matters: evolution of old world Salvia (Lamiaceae) in Africa. *Ann. Bot.* 114, 61–83.

Will, M., and Claßen-Bockhoff, R. (2017). Time to split Salvia s.l. (Lamiaceae) – new insights from Old World Salvia phylogeny. *Mol. Phylogenet. Evol.* 109, 33–58.

Yu, Y., and Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16:S10. doi: 10.1186/1471-2164-16-S10-S10

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

Zhao, F., Chen, Y. P., Salmaki, Y., Drew, B. T., Wilson, T. C., Scheen, A. C., et al. (2021). An updated tribal classification of Lamiaceae based on plastome phylogenomics. *BMC Biol.* 19:2. doi: 10.1186/s12915-020-00931-z

Zhao, F., Drew, B. T., Chen, Y. P., Hu, G. X., Li, B., and Xiang, C. L. (2020). The chloroplast genome of Salvia: Genomic characterization and phylogenetic analysis. *Intl. J. Plant Sci.* 181, 812–830.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 06 January 2022
doi: 10.3389/fpls.2021.720171

Mahonia vs. *Berberis* Unloaded: Generic Delimitation and Infrafamilial Classification of Berberidaceae Based on Plastid Phylogenomics

*Chia-Lun Hsieh[1], Chih-Chieh Yu[1,2], Yu-Lan Huang[1] and Kuo-Fang Chung[1]**

[1] *Biodiversity Research Center, Academia Sinica, Taipei, Taiwan,* [2] *School of Forestry and Resource Conservation, National Taiwan University, Taipei, Taiwan*

The early-diverging eudicot family Berberidaceae is composed of a morphologically diverse assemblage of disjunctly distributed genera long praised for their great horticultural and medicinal values. However, despite century-long studies, generic delimitation of Berberidaceae remains controversial and its tribal classification has never been formally proposed under a rigorous phylogenetic context. Currently, the number of accepted genera in Berberidaceae ranges consecutively from 13 to 19, depending on whether to define *Berberis*, *Jeffersonia*, and *Podophyllum* broadly, or to segregate these three genera further and recognize *Alloberberis*, *Mahonia*, and *Moranothamnus*, *Plagiorhegma*, and *Dysosma*, *Diphylleia*, and *Sinopodophyllum*, respectively. To resolve Berberidaceae's taxonomic disputes, we newly assembled 23 plastomes and, together with 85 plastomes from the GenBank, completed the generic sampling of the family. With 4 problematic and 14 redundant plastome sequences excluded, robust phylogenomic relationships were reconstructed based on 93 plastomes representing all 19 genera of Berberidaceae and three outgroups. Maximum likelihood phylogenomic relationships corroborated with divergence time estimation support the recognition of three subfamilies Berberidoideae, Nandinoideae, and Podophylloideae, with tribes Berberideae and Ranzanieae, Leonticeae and Nandineae, and Podophylleae, Achlydeae, Bongardieae *tr. nov.*, Epimedieae, and Jeffersonieae *tr. nov*. in the former three subfamilies, respectively. By applying specifically stated criteria, our phylogenomic data also support the classification of 19 genera, recognizing *Alloberberis*, *Mahonia*, and *Moranothamnus*, *Plagiorhegma*, and *Diphylleia*, *Dysosma*, and *Sinopodophyllum* that are morphologically and evolutionarily distinct from *Berberis*, *Jeffersonia*, and *Podophyllum*, respectively. Comparison of plastome structures across Berberidaceae confirms inverted repeat expansion in the tribe Berberideae and reveals substantial length variation in *accD* gene caused by repeated sequences in Berberidoideae. Comparison of plastome tree with previous studies and nuclear ribosomal DNA (nrDNA) phylogeny also reveals considerable conflicts at different phylogenetic levels, suggesting that incomplete lineage sorting and/or hybridization had occurred throughout the evolutionary history of Berberidaceae and that *Alloberberis* and *Moranothamnus* could have resulted from reciprocal hybridization between *Berberis* and *Mahonia* in ancient times prior to the radiations of the latter two genera.

Keywords: *accD* length variation, cytonuclear discordance, IR expansion, molecular dating, tribal classification

OPEN ACCESS

Edited by:
Susann Wicke, Humboldt University of Berlin, Germany

Reviewed by:
Ana Otero, Field Museum of Natural History, United States
Ross McCauley, Fort Lewis College, United States
Diego F. Morales-Briones, Ludwig Maximilian University of Munich, Germany

***Correspondence:**
Kuo-Fang Chung
bochung@gate.sinica.edu.tw

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *03 June 2021*
Accepted: *15 October 2021*
Published: *06 January 2022*

Citation:
Hsieh C-L, Yu C-C, Huang Y-L and Chung K-F (2022) Mahonia vs. Berberis Unloaded: Generic Delimitation and Infrafamilial Classification of Berberidaceae Based on Plastid Phylogenomics. Front. Plant Sci. 12:720171. doi: 10.3389/fpls.2021.720171

INTRODUCTION

The early-diverging eudicot family Berberidaceae is composed of a morphological diverse assemblage of genera (**Figure 1**) long praised for their great horticultural (Ahrendt, 1961; Stearn, 2002) and medicinal values (Peng et al., 2006; Hao, 2018). Although more than 85% of the ca. 700 species of Berberidaceae (Christenhusz and Byng, 2016) are woody shrubs (Yu and Chung, 2017), at the generic level, the family is predominantly represented by mono- and oligotypic temperate herbaceous genera known for several classic examples of biogeographic disjunctions (Liu et al., 2002; Wang et al., 2007; Zhang et al., 2007; Sun et al., 2018).

In the Northern Hemisphere, Berberidaceae exhibits seven intercontinental disjunctions: the East Asian (EA) and western North American (WNA) disjunctions in *Achlys* (Fukuda, 1967) and *Mahonia* (Yu and Chung, 2017; Chen et al., 2020), the Eurasian *Epimedium* and its WNA disjunct sister genus *Vancouveria* (Stearn, 1938; Zhang et al., 2007), the EA and eastern North American (ENA) disjunctions in *Diphylleia* (Ying et al., 1984) and *Caulophyllum* (Loconte and Blackwell, 1985), and the EA monotypic genera *Sinopodophyllum* (Ying, 1979) and *Plagiorhegma* (Hutchinson, 1920) and their respective disjunct ENA sister genera *Podophyllum* and *Jeffersonia* (Wang et al., 2007). Because of great economic, ecological, and taxonomic interests, Berberidaceae has been studied extensively in seedling morphology (Terabayashi, 1985c), floral morphology (Terabayashi, 1985a; Brückner, 2000), embryology (Sastri, 1969), serology (Jensen, 1973), palynology (Zhang et al., 2017), wood anatomy (Carlquist, 1995), and chromosome cytology (Kuroki, 1970; Adhikari et al., 2014; Huang et al., 2018; Wang et al., 2020).

Historically, however, owing to the heterogeneous composition of the family that is "*held together more by a linkage of characteristics than by possession of any set of diagnostic features* (Meacham, 1980)," Berberidaceae had been variously segregated into smaller families including Nandinaceae, Leonticaceae, Podophyllaceae, and Ranzaniaceae (e.g., Janchen, 1949; Airy Shaw, 1973; Hutchinson, 1973; Wu et al., 2003; Takhtajan, 2009; Lu and Tang, 2020), and/or classified into different infrafamilial taxa including subfamilies (i.e., Berberidoideae, Epimedioideae, Leonticoideae, Nandinoideae, and Podophylloideae) and tribes (i.e., Achlydeae, Berberideae, Bongardieae, Epimedieae, Leonticeae, Podophylleae, and Ranzanieae) (**Table 1**). Additionally, as stated in the popular encyclopedia "Flowering Plant Families of the World" that Berberidaceae contains "12 to 16" genera (Heywood et al., 2007), generic delimitation of the family has long been disputed and thus the number of its recognized genera varies greatly (**Table 1** and **Supplementary Table 1**). Indeed, there seems no consensus regarding whether to adopt a broadly defined *Berberis* (e.g., Sun et al., 2018; Kreuzer et al., 2019), *Jeffersonia* (e.g., Sun et al., 2016), and *Podophyllum* (e.g., Shaw, 2002; Christenhusz et al., 2018), or to recognize *Alloberberis*, *Mahonia*, and *Moranothamnus*, *Plagiorhegma*, and *Diphylleia*, *Dysosma*, and *Sinopodophyllum* as distinct genera separated from the former three genera (**Supplementary Table 1**). In particular, whether *Mahonia* (i.e., the compound-leaved *Berberis*) should be synonymized under a broad sense *Berberis* (*Berberis s.l.*) has been debated for more than two centuries (Fedde, 1901; Moran, 1982; Kim et al., 2004b; Adhikari et al., 2015; Yu and Chung, 2017). Please refer to Ahrendt (1961) and Yu and Chung (2017) for more details about the *Berberis* vs. *Mahonia* debates.

To resolve Berberidaceae's taxonomic controversies, early molecular studies using nuclear glyceraldehyde-3-phosphate dehydrogenase gene (Adachi et al., 1995) and chloroplast *rbcL* gene and restriction site (Kim and Jansen, 1995) both showed that *Nandina* should be included within the family. Subsequent molecular phylogenetic studies (Kim and Jansen, 1996; Kim et al., 2004a; Wang et al., 2007) revealed three clades within Berberidaceae, resulting in the circumscription of three subfamilies corresponding to three chromosome groups (Wang et al., 2009): Berberidoideae ($x = 7$), Podophylloideae ($x = 6$), and Nandinoideae ($x = 8$ and $x = 10$). Except for Lu and Tang (2020), Wang et al.'s (2009) subfamilial classification of Berberidaceae has been widely followed (**Table 1**). Subsequent historical biogeographic analyses based on molecular phylogenetic data also indicate that the Bering Land Bridge had functioned as a crucial pathway for the intercontinental disjunctions (Wen et al., 2010). Based on the internal transcribed spacer (ITS), Kim et al. (2004b) showed that *Mahonia* is paraphyletic, with *Mahonia* sect. *Horridae* sister to the simple-leaved *Berberis* (i.e., *Berberis s.s.*). More recently, based on the combined analysis of ITS and chloroplast *ndhF* gene sequences, Adhikari et al. (2015) further showed that Sect. *Horridae* is polyphyletic, together with Kim et al. (2004b) arguing for a broadly circumscribed *Berberis* (i.e., *Berberis s.l.*) that includes the compound-leaved *Mahonia*.

However, both Kim et al. (2004b) and Adhikari et al. (2015) suffered from issues including inadequate taxon sampling, problematic outgroup rooting, inclusion of poor-quality DNA sequences from GenBank, and taxon misidentification, undermining their taxonomic conclusion (Yu and Chung, 2017). To resolve the *Mahonia* vs. *Berberis* debate that has been lasting for more than two centuries, Yu and Chung (2017) expanded and verified taxon sampling of *Mahonia* and included *Berberis claireae*, a unique spineless Baja California endemic species with unifoliolate to 7-foliolate compound leaves (Moran, 1982) that had never been sampled previously. Based on ITS and four cpDNA markers, Yu and Chung's (2017) phylogenetic analyses of *Berberis s.l.* revealed four strongly supported clades, *Berberis s.s.*, *B. claireae*, core *Mahonia*, and *Mahonia* sect. *Horridae*. Because these four clades are ecologically and morphologically distinct and evolutionarily comparable to other genera of Berberidaceae, Yu and Chung (2017) proposed a new classification that recognizes these four clades as genera: *Alloberberis* (≡ *Mahonia* sect. *Horridae*), *Berberis* (≡ *Berberis s.s.*), *Mahonia* (≡ core *Mahonia*), and, *Moranothamnus* (≡ *B. claireae*), "reloading" the two-century long "*Mahonia* vs. *Berberis*" debate (see cover of the journal *Taxon* 66(6); doi.org/10.1002/tax.666001).

However, debates on generic concepts of Berberidaceae are not restricted to *Mahonia* vs. *Berberis*. Neither do controversies end with phylogenetic and phylogenomic data. In both Kim et al. (2004a) and Wang et al. (2007), Berberidaceae were regarded as having 17 genera; however, in the first molecular-based formal infrafamilial classification of Berberidaceae, Wang et al. (2009)

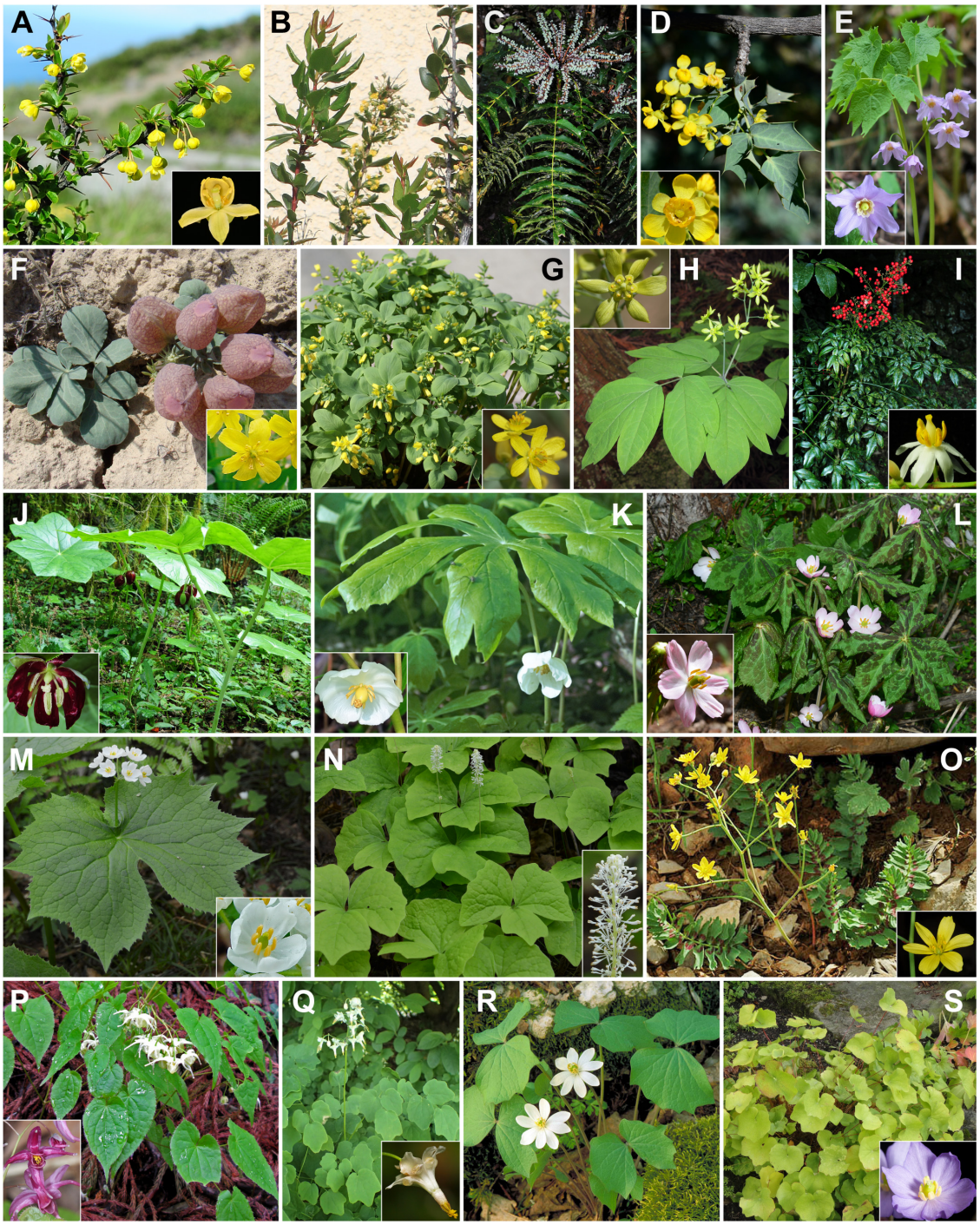


FIGURE 1 | Morphological diversity in Berberidaceae. **(A)** *Berberis morrisonensis* and *B. mingetsensis* (flower). **(B)** *Moranothamnus claireae*, courtesy of Bart O'Brien. **(C)** *Mahonia oiwakensis*. **(D)** *Alloberberis fremontii* and flower photo of *A. nevinii* by Stan Shebs/CC BY-SA 3.0. **(E)** *Ranzania japonica*, courtesy of Takuro Ito, and flower photo by Qwert1234/CC BY-SA 3.0. **(F)** *Leontice incerta*, photo by Yuriy Danilevsky/CC BY-SA 3.0 and *L. leotopetalum* (flowers), photo by Averater/CC BY-SA 3.0. **(G)** *Gymnospermium altaicum*, photos by Ettrig/CC BY-SA 4.0. **(H)** *Caulophyllum robustum*, photo by Qwert1234/CC BY-SA 3.0, flower photo by Alpsdake/CC BY-SA 4.0. **(I)** *Nandina domestica*. **(J)** *Dysosma pleiantha*. **(K)** *Podophyllum peltatum*, photo by WilderAddict/CC BY-SA 4.0, flower photo by Nicholas A. Tonelli/CC BY 2.0. **(L)** *Sinopodophyllum hexandrum*, courtesy of Mu-Tan Hsieh. **(M)** *Diphylleia grayi*, courtesy of Takuro Ito, and flower photo by yamatsu/CC0 1.0. **(N)** *Achlys triphylla*, courtesy of Takuro Ito. **(O)** *Bongardia chrysogonum*, photos by Ori Fragman-Sapir/CC BY 3.0. **(P)** *Epimedium koreanum*, photo by Qwert1234/CC BY-SA 3.0 and flower photo of *E. grandiflorum* var. *thunbergianum* by Alpsdake/CC BY-SA 3.0. **(Q)** *Vancouveria hexandra*, photo by Krzysztof Ziarnek, Kenraiz/CC BY-SA 4.0, and flower photo by Walter Siegmund/CC BY-SA 3.0. **(R)** *Jeffersonia diphylla*, photo by Barnes Dr. Thomas G, U.S. Fish and Wildlife Service. **(S)** *Plagiorhegma dubium*, photo by Daderot/CC0 1.0 and flower photo by sunoochi/CC BY 2.0.

TABLE 1 | Different classification systems proposed for Berberidaceae.

Present study	Janchen (1949)	Airy Shaw (1973)	Hutchinson (1973)	Meacham (1980)[1]	Terabayashi (1985b)	Loconte and Estes (1989)
Berberidaceae (N^2 = 19) Berberidoideae Berberideae *Alloberberis*, *Berberis*, *Mahonia*, *Moranothamnus* Ranzanieae *Ranzania* Nandinoideae Leonticeae *Caulophyllum*, *Gymnospermium*, *Leontice* Nandineae *Nandina* Podophylloideae Achlydeae *Achlys* Bongardieae *tr. nov.* *Bongardia* Epimedieae *Epimedium*, *Vancouveria* Jeffersonieae *tr. nov.* *Jeffersonia*, *Plagiorhegma* Podophylleae *Diphylleia*, *Dysosma*, *Podophyllum*, *Sinopodophyllum*	**Berberidaceae** (*N* = 15) Berberidoideae Berberideae Berberidinae *Berberis*, *Mahonia* Ranzaniinae *Ranzania* Epimedieae Epimediinae *Bongardia*, *Caulophyllum*, *Epimedium*, *Gymnospermium*, *Jeffersonia*, *Leontice*, *Plagiorhegma*, *Vancouveria* Achlyinae *Achlys* Podophylloideae Podophylleae Podophyllinae *Dysosma*, *Podophyllum* Diphylleiinae *Diphylleia* **Nandinaceae** (*N* = 1) *Nandina*	**Berberidaceae** (*N* = 4) *Berberis*, *Epimedium*, *Mahonia*, *Vancouveria* **Leonticaceae** (*N* = 4) *Bongardia*, *Caulophyllum*, *Gymnospermium*, *Leontice* **Nandinaceae** (*N* = 1) *Nandina* **Podophyllaceae** (*N* = 7) *Achlys*, *Diphylleia*, *Dysosma*, *Jeffersonia*, *Plagiorhegma*, *Podophyllum*, *Ranzania*	**Berberidaceae** (*N* = 2) *Berberis*, *Mahonia* **Nandinaceae** (*N* = 1) *Nandina* **Podophyllaceae** (*N* = 13) *Achlys*, *Bongardia*, *Caulophyllum*, *Diphylleia*, *Dysosma*, *Epimedium*, *Gymnospermium*, *Jeffersonia*, *Leontice*, *Plagiorhegma*, *Podophyllum*, *Ranzania*, *Vancouveria*	**Berberidaceae** (*N* = 15) Berberidoideae *Berberis*, *Mahonia*, *Ranzania* Podophylloideae *Diphylleia*, *Dysosma*, *Podophyllum* Epimedioideae *Achlys*, *Epimedium*, *Jeffersonia*, *Plagiorhegma*, *Vancouveria* Leonticoideae *Bongardia*, *Caulophyllum*, *Gymnospermium*, *Leontice* **Nandinaceae** (*N* = 1) *Nandina*	**Berberidaceae** (*N* = 16) Berberidoideae Berberideae *Berberis*, *Mahonia* Ranzanieae *Ranzania* Epimedieae Epimediinae *Achlys*, *Epimedium*, *Jeffersonia*, *Plagiorhegma*, *Vancouveria* Leonticinae *Bongardia*, *Caulophyllum*, *Leontice*, *Gymnospermium* Podophylleae *Diphylleia*, *Dysosma*, *Podophyllum* Nandinoideae *Nandina*	**Berberidaceae** (*N* = 17) Berberidoideae Berberideae Berberidinae *Berberis*, *Mahonia*, *Ranzania* Epimediinae *Achlys*, *Bongardia*, *Diphylleia*, *Dysosma*, *Epimedium*, *Jeffersonia*, *Plagiorhegma*, *Podophyllum*, *Sinopodophyllum*, *Vancouveria* Leonticeae *Caulophyllum*, *Leontice*, *Gymnospermium* Nandinoideae *Nandina*

Thorne (1992)	Loconte (1993) and Loconte et al. (1995)	Takhtajan (1997) and Takhtajan (2009)	Thorne (2000) and Thorne and Reveal (2007)	Wang et al. (2009)	Wu et al. (2003) and Lu and Tang (2020)
Berberidaceae (*N* = 16) Berberidoideae *Berberis*, *Mahonia*, *Ranzania* Leonticoideae *Caulophyllum*, *Leontice*, *Gymnospermium* Epimedioideae *Achlys*, *Bongardia*, *Dysosma*, *Diphylleia*, *Epimedium*, *Jeffersonia*, *Plagiorhegma*, *Podophyllum* (+*Sinopodophyllum*), *Vancouveria* Nandinoideae *Nandina*	**Berberidaceae** (*N* = 15) Berberidoideae Berberideae Berberidinae *Berberis*, *Mahonia*, *Ranzania* Epimediinae *Achlys*, *Bongardia*, *Dysosma*, *Epimedium*, *Jeffersonia* (+*Plagiorhegma*), *Vancouveria*, *Podophyllum* (+*Sinopodophyllum*) Leonticeae *Caulophyllum*, *Diphylleia*, *Gymnospermium*, *Leontice* Nandinoideae *Nandina*	**Berberidaceae** (*N* = 2) *Berberis*, *Mahonia* **Ranzaniaceae** (*N* = 1) *Ranzania* **Podophyllaceae** (*N* = 12) Leonticoideae *Caulophyllum*, *Gymnospermium*, *Leontice* Epimedioideae Epimedieae *Epimedium*, *Vancouveria*, *Jeffersonia*, *Plagiorhegma* Achlydeae *Achlys* Bongardieae *Bongardia* Podophylloideae *Diphylleia*, *Dysosma*, *Podophyllum* (+*Sinopodophyllum*) **Nandinaceae** (*N* = 1) *Nandina*	**Berberidaceae** (*N* = 13) Berberidoideae *Berberis* (+*Mahonia*), *Ranzania* Leonticoideae *Caulophyllum*, *Leontice*, *Gymnospermium* Podophylloideae *Achlys*, *Bongardia*, *Dysosma* (+*Diphylleia*?), *Epimedium*, *Jeffersonia* (+*Plagiorhegma*), *Podophyllum* (+*Sinopodophyllum*), *Vancouveria* Nandinoideae *Nandina*	**Berberidaceae** (*N* = 16) Berberidoideae *Berberis*, *Mahonia*, *Ranzania* Podophylloideae *Achlys*, *Diphylleia*, *Dysosma*, *Podophyllum*, *Sinopodophyllum*, *Bongardia*, *Epimedium*, *Vancouveria*, *Jeffersonia*, *Plagiorhegma* Nandinoideae *Caulophyllum*, *Gymnospermium* (+*Leontice*), *Nandina*	**Berberidaceae** (*N* = 3) Berberideae *Berberis*, *Mahonia* Ranzanieae *Ranzania* **Leonticaceae** (*N* = 3) *Caulophyllum*, *Gymnospermium*, *Leontice* **Podophyllaceae** (*N* = 10) Epimedioideae Epimedieae *Epimedium*, *Vancouveria*, *Jeffersonia*, *Plagiorhegma* Achlydeae *Achlys* Bongardieae *Bongardia* Podophylloideae *Diphylleia*, *Dysosma*, *Podophyllum*, *Sinopodophyllum* **Nandinaceae** (*N* = 1) *Nandina*

[1]*Meacham's (1980) analysis supports the recognition of four "subfamilial taxa" without formal taxonomic treatment; the four subfamilies presented here are added based on taxonomic priority.* [2]*N, the number of genera.*

sampled "*all 16 genera of Berberidaceae*," neglecting *Leontice* L. that had never previously been synonymized (**Table 1**). In a recent phylogenomic study using plastome sequences, Sun et al. (2018) recognized 18 genera in Berberidaceae, accepting Yu and Chung's (2017) new genera *Alloberberis* and *Moranothamnus* and yet subsuming *Plagiorhegma* under *Jeffersonia* (**Table 2**). However, in a subsequent study aiming to develop clade-specific DNA barcodes of *Berberis* using plastome sequences, sampled *Alloberberis nevinii*, *Mahonia nervosa*, and *M. polyodonta* were all treated as *Berberis s.l.* (Kreuzer et al., 2019). The flux of Berberidaceae's generic delimitation is also manifested across major biodiversity databases and online resources (**Supplementary Table 1**). Nevertheless, Yu and Chung's (2017) classification has been taken by taxonomic (Colin et al., 2021), floristic (Galasso et al., 2018), paleobotanical (Doweld, 2018), and biogeographic (Chen et al., 2020) studies.

The disparity on generic concepts across different studies and online resources (**Supplementary Table 1**) illustrates the lack of consensus on precise and objective criteria for generic delimitation (Humphreys and Linder, 2009) in Berberidaceae. Indeed, most of the abovementioned studies and online resources did not specify references or explicitly state reasons for their adoption of a particular generic treatment. To achieve an objective generic delimitation of *Berberis s.l.*, Yu and Chung (2017) followed strictly five criteria advocated by Backlund and Bremer (1998), Linder et al. (2010), and Heenan and Smissen (2013) to delimit *Berberis s.l.*: (1) prioritizing primary (i.e., family, genus, and species) over secondary ranks (i.e., subgenus, section, etc.), (2) maximizing phylogenetic information and reducing redundancy in a classification, (3) recognizing evolutionarily equivalent (i.e., clade age, phylogenetic distance, and morphology) groups as the same rank, (4) delimiting genus that is morphologically, ecologically, and geographically homogenous, and (5) taking into account the full taxonomic history of the group and minimizing name changes to maintain nomenclatural stability. However, such objective generic delimitation has not been applied to other genera of Berberidaceae.

In recent years, rapid advances in high-throughput sequencing technology have made plastome sequences accessible for resolving recalcitrant phylogenetic relationships not attainable previously using Sanger sequences (Wicke and Schneeweiss, 2015; Tonti-Filippini et al., 2017; Gitzendanner et al., 2018). Several phylogenomic studies of Berberidaceae have been conducted using whole plastome sequences (Zhang et al., 2016; Sun et al., 2018; Ye et al., 2018; Kreuzer et al., 2019); however, no phylogenomic studies have yet sampled all 19 genera and covered adequate infrageneric diversity needed to resolve the taxonomic controversies. In this study, we report 23 newly assembled plastome sequences that complete the generic sampling of Berberidaceae. By implementing explicit criteria of generic delimitation, an infrafamilial classification representing monophyletic subdivisions of Berberidaceae is proposed, aiming to settle the taxonomic controversies and debates that have been fraught for centuries.

MATERIALS AND METHODS

Classification Adopted

For clarity, the classification of 19 genera in Berberidaceae (**Table 1**) that recognizes *Alloberberis*, *Berberis*, *Mahonia*, and *Moranothamnus* (Yu and Chung, 2017), *Jeffersonia* and *Plagiorhegma* (Hutchinson, 1920), and *Diphylleia*, *Dysosma*, *Podophyllum*, and *Sinopodophyllum* (Wang et al., 2009) as opposed to the broadly defined *Berberis s.l.*, *Jeffersonia s.l.*, and *Podophyllum s.l.*, respectively, is followed in all subsequent discussion unless otherwise stated.

Taxon Sampling

A total of 85 plastomes representing 60 species and two additional varieties in 17 genera of Berberidaceae available (accessed 25 March 2021) on GenBank were downloaded (**Supplementary Table 2**). To complete generic ($N = 19$) and infrageneric sampling of Berberidaceae, 23 species of Berberideae, including 3 species of *Alloberberis*, the monotypic *Moranothamnus*, 8 species of *Berberis* (7 species of Group Septentrionales and 1 species of Group Australes), and 11 species of *Mahonia* (5 species of Group Orientales and 6 of Group Occidentales), were sampled (**Supplementary Table 2**) for plastome assembly. Although we only sampled 11 species and two additional varieties of the ca. 500 species of *Berberis* and 11 of the ca. 100 species of *Mahonia*, our sampling is geographically and phylogenetically sufficient (Yu and Chung, 2017; Yu, 2018) to address issues of generic circumscription in Berberidaceae. Based on recent studies (e.g., Lane et al., 2018), plastomes of *Ranunculus macrantha* (Ranunculaceae), *Stephania japonica* (Menispermaceae), and *Akebia quinata* (Lardizabalaceae) were also downloaded from GenBank as outgroups (**Supplementary Table 2**).

DNA Extraction and Next-Generation Sequencing

CTAB method (Doyle and Doyle, 1987) was used to extract total genomic DNA from silica-dried and herbarium leaf materials. The DNA concentration was quantified by Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, United States). The DNAs were sent to the Genomic Core Lab of Institute of Molecular Biology, Academia Sinica for library preparation using KAPA LTP Library Preparation Kits (KAPA Biosystems, Wilmington, MA, United States), and for whole genome shotgun (WGS) sequencing using Illumina NextSeq 500 (Illumina Inc., San Diego, CA, United States) with pair-end mode, read length = 150 bp, and insert size = ca. 300 bp.

Plastome Assembly and Annotation

The quality of raw reads was assessed by FastQC v.0.11.9 (Andrews, 2010). Reads were trimmed using Trimmomatic v.0.39 (Bolger et al., 2014) with the setting "LEADING:25 TRAILING:25 SLIDINGWINDOW:4:20 CROP:149 MINLEN:100." The *de novo* assembly of the plastome was performed by GetOrganelle v.1.7.5 (Jin et al., 2020) with the setting of "-R 10 -t 3 -w 0.8 -k 37,55,65,85,105,127,131 -F embplant_pt –reduce-reads-for-coverage inf," using *Berberis amurensis* (GenBank accession:

TABLE 2 | Summary of the plastome and nrDNA assembly data.

		Plastome								nrDNA			
Species	**# all reads[1]**	**NCBI accession**	**Length (bp)**	**LSC (bp)**	**SSC (bp)**	**IR (bp)**	**%GC**	**Av. cov. (×)**	**Cov. SD**	**NCBI accession**	**Length (bp)**	**Av. cov. (×)**	**Cov. SD**
Alloberberis fremontii	8,924,456	MT335778	165,871	73,262	18,779	36,915	38.1	496.2	141.8	MW545966	7300	1454	1096.3
A. higginsiae	10,224,582	MT335779	165,883	73,279	18,788	36,908	38.1	902.0	269.6	MW545967	7300	1823.3	1367
A. trifoliolata	10,646,702	MT335780	164,553	72,349	18,738	36,733	38.1	101.5	31.1	MW545968	6906	1893.0	2156.7
Berberis dictyophylla	9,547,576	MT335782	166,036	73,449	18,611	36,988	38.1	57.0	16.3	MW545974	7198	987.0	258.3
B. hayatana	10,975,726	MT335783	168,208	73,245	16,277	39,343	38.0	333.7	66.5	MW545975	7220	1171.9	304.7
B. kawakamii	9,066,338	MT335784	167,658	73,294	16,194	39,085	38.1	212.7	45.7	MW545976	7221	933.9	243.6
B. morrisonensis	11,053,602	MT335785	166,145	73,490	18,623	37,016	38.1	331.2	74.2	MW545979	7198	1476.5	280.6
B. nantoensis	9,136,400	MT335806	167,898	73,296	16,270	39,166	38.0	267.7	52.5	MW545980	7220	944.5	179
B. pruinosa	9,574,826	MT335786	165,455	73,348	18,573	36,767	38.1	95.4	23.4	MW545982	7260	1217.0	318.5
B. saxicola	9,458,418	MT335787	166,172	73,606	18,692	36,937	38.1	128.2	34.2	MW545984	6843	940.1	268.4
B. vulgaris	10,088,158	MT335788	166,150	73,460	18,660	37,015	38.0	324.4	71.3	MW545987	7192	2063.8	561.9
Mahonia aquifolium	9,557,402	MT335789	165,517	73,149	18,758	36,805	38.1	546.9	173.8	MW545988	7110	1372.8	597.8
M. chochoco	11,204,800	MT335790	165,367	73,301	18,682	36,692	38.1	251.4	100.6	MW545989	7322	750.3	702
M. dictyota	10,890,484	MT335791	165,495	73,065	18,824	36,803	38.1	172.5	45.2	MW545990	7110	955.8	240.2
M. fortunei	10,301,560	MT335792	165,654	73,669	18,623	36,681	38.0	133.2	30.5	MW545991	7165	866.4	278.8
M. harrisoniana	10,575,142	MT335793	165,367	73,095	18,822	36,725	38.1	967.1	263.1	MW545992	7110	920.5	367.6
M. japonica	9,972,444	MT335794	164,827	73,253	18,634	36,470	38.2	484.4	97.6	MW545993	7313	2547.7	849.7
M. lanceolata	9,623,936	MT335795	165,796	72,886	18,744	37,083	38.0	297.7	66.3	MW545994	7168	1294.6	494.2
M. nervosa	11,134,750	MT335796	165,707	73,128	18,825	36,877	38.1	119.4	30.5	MW545995	7346	1354.6	500.5
M. oiwakensis	7,600,014	MT335797	165,021	73,260	18,649	36,556	38.1	609.3	115.9	MW545996	7324	853.7	329.3
M. pallida	11,120,770	MT335798	165,707	72,782	18,717	37,104	38.0	298.5	67.8	MW545997	7202	1373.4	505.4
M. tikushiensis	8,869,818	MT335799	164,876	73,273	18,713	36,445	38.1	226.5	11.3	MW545998	7313	1062.6	378
Moranothamnus claireae	10,834,754	MT335800	165,706	73,324	18,932	36,725	38.1	95.9	25.4	MW545999	7232	1176.0	686.9

LSC, large single copy; SSC, small single copy; IR, inverted repeat; %GC, GC content percentage; av. cov., average coverage; #, the number of; cp, chloroplast; SD, standard deviation.
[1]Number of all the trimmed reads of the sample.

KM057374) as a reference for assembly. The resulting sequences generated by GetOrganelle were imported into Geneious Prime (Biomatters Ltd., Auckland, New Zealand) (Kearse et al., 2012) for validation and/or final assembly completion. For samples not assembled into a complete plastome using GetOrganelle, the "Map to Reference" function with "High Sensitivity" and default setting of Geneious was implemented to generate the draft genome, using the consensus of the mapping file to temporarily fill the "unassembled regions." The "unassembled regions" were corrected by mapping the trimmed reads to the draft genome using the "Map to Reference" function with "Medium-Low Sensitivity" and default setting in order to complete the assembly. All complete plastome sequences were further verified by read mapping.

Newly assembled plastomes were annotated by transferring the annotations of published Berberidaceae plastomes to the newly sequenced ones under the alignment generated by MAFFT v.7.388 (Katoh and Standley, 2013) launched in Geneious. The presence of start and stop codons of each protein-coding gene was checked and adjusted manually. Genes with any premature stop codon that might interrupt translations from half of the original reading frame were annotated as pseudogenes. The correct length and identity of tRNA genes were further confirmed using the web server tRNAscan-SE 2.0 (Lowe and Chan, 2016). The boundaries of IRs were annotated by GeSeq (Tillich et al., 2017) and manually checked with self-dot plots under Geneious. Plastome maps were drawn using OGDRAW (Greiner et al., 2019).

Plastome Phylogenetic Analyses

Our initial matrix comprised 111 plastomes, including 108 of Berberidaceae (81 species and 2 additional varieties in 19 genera) and 3 outgroups (**Supplementary Table 2**). The sequence MG593045 (*Dysosma delavayi*) was excluded because high sequence variation was detected between its two inverted repeats (IRs). Of the remaining 80 species of Berberidaceae, 18 species were represented by multiple sequences. To lessen computational loading, we conducted a preliminary maximum likelihood (ML) analysis of the 110 sequences using IQ-TREE v.1.6.12 (Nguyen et al., 2015). Based on the preliminary ML tree (**Supplementary Figure 1**), 14 redundant and 3 problematic sequences were further excluded (see section "Results"), leaving a total of 93 plastomes representing 80 species and 2 additional varieties in all 19 genera of Berberidaceae and 3 outgroups for subsequent analyses.

Prior to phylogenetic analyses, IRB was removed. To accommodate substitution rate heterogeneity across plastomes, sequences were partitioned by the four gene categories [i.e., coding sequences (CDSs) of protein-coding genes, introns, RNA (tRNA and rRNA) genes, and intergenic spacers (IGSs)] as well as codon position of CDS. Each category was extracted, concatenated, and aligned individually by MAFFT using Geneious. For CDS, after excluding pseudogenes and partially duplicated genes, the remaining 76 genes (**Supplementary Table 3**) were concatenated and aligned using the "Translation Align" function based on bacterial genetic codes implemented by MAFFT under Geneious, with manual adjustments. The final concatenated alignment contains six partitions (i.e., plastid partition scheme): CDS1, CDS2, CDS3, introns, RNA genes, and IGS. Sites with more than 97% gaps were excluded using "Mask Alignment" function in Geneious. The number and proportion of parsimony informative sites of the concatenated plastome alignment were calculated by AMAS (Borowiec, 2016).

IQ-TREE was used with the "-m MFP+MERGE -bb 5000" option to conduct the following analyses: (1) searching for the best-fit partition scheme, (2) determining the best-fit nucleotide model for each partition by ModelFinder (Kalyaanamoorthy et al., 2017), and (3) reconstructing phylogenies based on ML method with 5000 replicates using ultrafast bootstrap approximation approach (Minh et al., 2013). The final tree with ultrafast bootstrap support (UFBS) values was visualized using FigTree v.1.4.2.[1]

Nuclear Ribosomal DNA Assembly and Analysis

Nuclear ribosomal DNA (nrDNA) sequences, spanning across partial external transcribed spacer (ETS), 18S rRNA gene, ITS 1, 5.8S rRNA gene, ITS2, 26S rRNA gene, and partial non-transcribed spacer (NTS) were assembled from raw reads of the 23 newly generated WGS sequencing using GetOrganelle with the setting of "-R 15 -t 10 -w 0.7 -k 37,69,85,115,127,131,135,139 -F embplant_nr –reduce-reads-for-coverage inf." Additionally, nrDNA were also assembled for *B. amurensis*, *B. koreana*, *B. weiningensis*, *Bongardia chrysogonum*, and *Podophyllum peltatum* from WGS sequencing reads downloaded from NCBI Sequence Read Archive (SRA) using NCBI SRA Toolkit v.2.1.11. The nrDNA of these samples were assembled by executing GetOrganelle with customized settings (**Supplementary Table 7**). All nrDNA were verified by read mapping with the same procedure as verifying plastome sequences.

The 28 nrDNA sequences were aligned and partitioned (i.e., partial ETS, 18S, ITS1, 5.8S, ITS2, 26S, and partial NTS) by MAFFT implemented in Geneious. We employed IQ-TREE with "-m MFP+MERGE -bb 5000" options to conduct the same analyses as the plastome dataset. Concurrently, a plastome tree including 28 species sampled for the nrDNA was generated by IQ-TREE using the same partition scheme and analytical settings of the 93-plastome dataset.

Divergence Times Estimation

For divergence times estimation, we kept only one sequence for each species to further reduce the computational time. As a result, the matrix including 83 plastome sequences of 80 species in 19 genera of Berberidaceae and 3 outgroups (**Supplementary Table 2**) was analyzed using BEAST v.2.6.0 (Bouckaert et al., 2019) on CIPRES Science Gateway v.3 (Miller et al., 2011). Parameters and priors of the input xml file were set *via* BEAUti launched in the software package of BEAST v.2.6.0. With IRB excluded, the analysis was performed with the plastid partition scheme, and the prior of site models were set according to the best-fit nucleotide models and partition scheme determined by ModelFinder in IQ-TREE with the options "-m TESTMERGEONLY -mset mrbayes." To accommodate rate

[1] tree.bio.ed.ac.uk/software/figtree/

heterogeneity across different Berberidaceae lineages (Yu and Chung, 2017; Sun et al., 2018), we used relaxed clock log normal as the prior of the clock model. The tree prior was set as a Yule model, and the remaining parameters followed default settings except for specifying three fossil calibration points to constrain the ages of three nodes. In Yu and Chung (2017), the age of the fossil *Leefructus mirus* (Sun et al., 2011; Wang et al., 2016) at 124.4 million years ago (Ma) was taken as the crown age of the Berberidaceae + Ranunculaceae clade. However, because of concern over the authenticity of the fossil of *L. mirus* (Zhou, 2014), three alternative fossils were used instead. First, the fossil of *Prototinomiscium vangerowii* dated back to the Turonian at ca. 91 Ma was assigned as the stem age of Menispermaceae (Anderson et al., 2005; Wang et al., 2012) with a lognormal distribution (mean = 92 in real space, SD = 0.06). Second, the fossil of *Mahonia simplex* from the Oligocene dated back to ca. 28.45 Ma (Huang et al., 2016) was designated as the crown age of *Mahonia* with the lognormal distribution (mean = 28.45 in real space, sigma = 0.1). Third, the fossil of *Alloberberis obliqua* from the Oligocene at ca. 35.55 Ma (MacGinitie, 1953; Doweld, 2018) was chosen as the crown age of *Alloberberis* with a lognormal distribution (mean = 35.55 in real space, SD = 0.05). We conducted two independent runs of Markov Chain Monte Carlo (MCMC), one with 400 million generations of MCMC and the other with 200 million generations. Both runs were sampled every 1000 steps for log files and every 50,000 steps for tree files. To evaluate the convergence of each parameter, the log file of each run was summarized and visualized by Tracer v.1.7.1 (Rambaut et al., 2018). The tree files were then combined by LogCombiner v.2.6.2 (launched in BEAST v.2.6.0) with the first 100 million trees discarded as burn-in for each run. Finally, we used TreeAnnotator v.2.6.0 (launched in the software package of BEAST v.2.6.0) to summarize the combined tree file into a maximum clade credibility tree with 95% highest posterior density (HPD) interval of age of each node calculated by mean heights, and visualized the tree using FigTree.

RESULTS

Plastome Features of Berberidoideae

All newly generated plastomes of Berberideae were assembled into circular molecules with sizes ranging from 164,553 (*Alloberberis trifoliolata*) to 168,208 bp (*Berberis hayatana*). The average coverages of the newly assembled plastomes ranged from 64× (*B. dictyophylla*) to 1127.8× (*Mahonia harrisoniana*) (**Table 2**). The GC contents vary only slightly (38.0–38.2%), and the genome structures are found to represent the typical quadripartite configuration (**Figure 2** and **Supplementary Figures 2–5**), consisting of a large single copy (LSC) ranging from 72,349 (*A. trifoliolata*) to 73,669 bp (*Mahonia fortunei*), a small single copy (SSC) ranging from 16,194 (*B. kawakamii*) to 18,932 bp (*Moranothamnus claireae*), and two IRs ranging from 36,445 (*M. tikushiensis*) to 39,343 bp (*B. hayatana*) (**Table 2**). Referring to early-diverging eudicots (Sun et al., 2016), both gene orders (**Figure 2** and **Supplementary Figures 2–5**) and gene contents (**Supplementary Table 3**) of the 23 newly assembled plastomes are consistent with the published plastome of *Mahonia bealei* (Ma et al., 2013), which has experienced significant IR expansions at IRB/LSC boundary from *rps19* into the spacer between *clpP* and *psbB* (**Supplementary Table 2** and **Supplementary Figure 6**). In addition to Berberideae, IR expansion was also detected in MG234280 of *Ranzania* (Wang et al., 2018) and *Epimedium ecalcaratum* (MN939634). On the other hand, IR contraction was found in MN371716 of *Epimedium brevicornu* (Zheng et al., 2019). However, both IR expansion and contraction are not previously known in *Epimedium*. Comparison of IR/SC boundaries across Berberidaceae is shown in **Supplementary Figure 6**. Additionally, as noted in Ma et al. (2013), *rpo*A gene is lost in all our newly sequenced plastomes of Berberideae. However, while Ma et al. (2013) reported that *ndhK* had degenerated into a pseudogene in *M. bealei*, *ndhK* gene does not contain any internal stop codon in all our newly assembled plastomes.

Together with all newly assembled plastomes, we also noticed a substantial length variation in *accD* genes in Berberidaceae (**Supplementary Figure 7**), especially in Berberidoideae (**Supplementary Figure 8**). All sampled plastomes of *Alloberberis* and *Mahonia* share a 216-bp deletion close to the 3′ end of the reading frame, with two additional deletions of 120 and 30 bp unique to the former genus (**Supplementary Figure 8**). However, the greatest sequence variation of *accD* locates in the central part of the gene. Visualizing the translation alignment revealed that the length variation in *accD* is featured by repeats composed of five amino acid sequences. In Berberidoideae, a total of 33 types of the amino acid repeats translated from 37 types of 15-bp DNA sequences were identified (**Supplementary Table 4**). The total number of these repeats in each species varies from 6 in *B. dictyophylla* to 27 in *Berberis aristata* (MN746308) and *B. saxicola*. Of the 33 amino acid repeats, R19 (120 copies) and R22 (87 copies) are the two most numerous copies, found in almost all plastomes of Berberideae (**Supplementary Table 5** and **Supplementary Figure 8**). Some repeats were detected in certain groups and thus appear to be clade specific. For example, R21 and R31 occur exclusively in Asian *Mahonia* clade (Group Orientales) except for *M. nervosa*, R10 is unique to *Mahonia*, and R8, R20, and R23 were found only in *Alloberberis* (**Supplementary Table 5** and **Supplementary Figure 8**).

Plastid Phylogenomic Analyses

Our preliminary ML analyses of the 110 plastome dataset indicated that, of the 18 species represented by multiple sequences, 10 species (*B. amurensis*, *M. fortunei*, *Ranzania japonica*, *Dysosma pleiantha*, *Diphylleia sinensis*, *Sinopodophyllum hexandrum*, *E. brevicornu*, *E. tianmenshanensis*, *E. wushanense*, and *Plagiorhegma dubium*) were recovered as monophyletic groups and two species (*Achlys triphylla* and *E. pseudowushanese*) were paraphyletic (**Supplementary Figure 1**). Of the two plastome sequences of *A. triphylla*, MG461315 was deleted for its poor sequence quality (Ye et al., 2018). For the two plastome sequences of *R. japonica*, MH423072 (Sun et al., 2018) was selected because MG234280

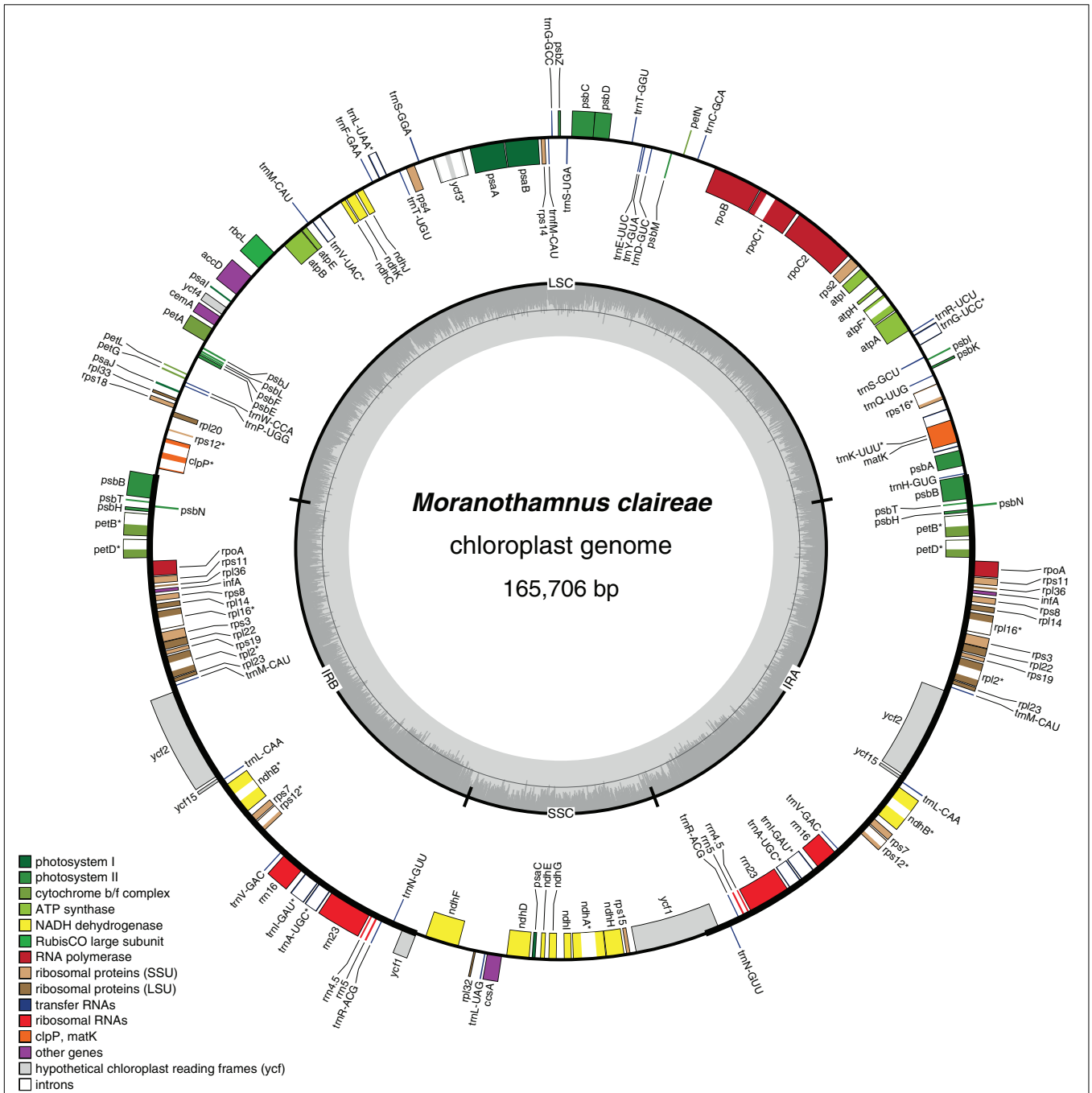


FIGURE 2 | Chloroplast genome map of *Moranothamnus claireae*. The genes drawn on the inner side of the outer circle are transcribed clockwise, and those on the outer side are transcribed counterclockwise. IRs are shown in bold lines in the outer circle. The darker gray areas of the inner circle indicate GC contents across the genome with lighter gray areas indicating AT contents. Genes belonging to different functional groups were shown in different color as in the legend.

(Wang et al., 2018) contains expanded IRs (**Supplementary Figure 6**) that was not reported by early chloroplast restriction site mapping study (Kim and Jansen, 1994). For the remaining eight monophyletic and one paraphyletic species, one plastome sequence was randomly selected for each species for subsequent analyses (**Supplementary Figure 1**). Seven species (*B. aristata*, *Mahonia oiwakensis*, *M. bealei*, *Dysosma versipellis*, *P. peltatum*, *Epimedium mikinorii*, and *E. sagittatum*) were shown to be polyphyletic (**Supplementary Figure 1**) and all their sequences were retained except for MG593052 (*P. peltatum*) that shares 99.19% of "% Identity" with *S. hexandrum* (KT445939) and yet only 87.2% with its conspecific sequence.

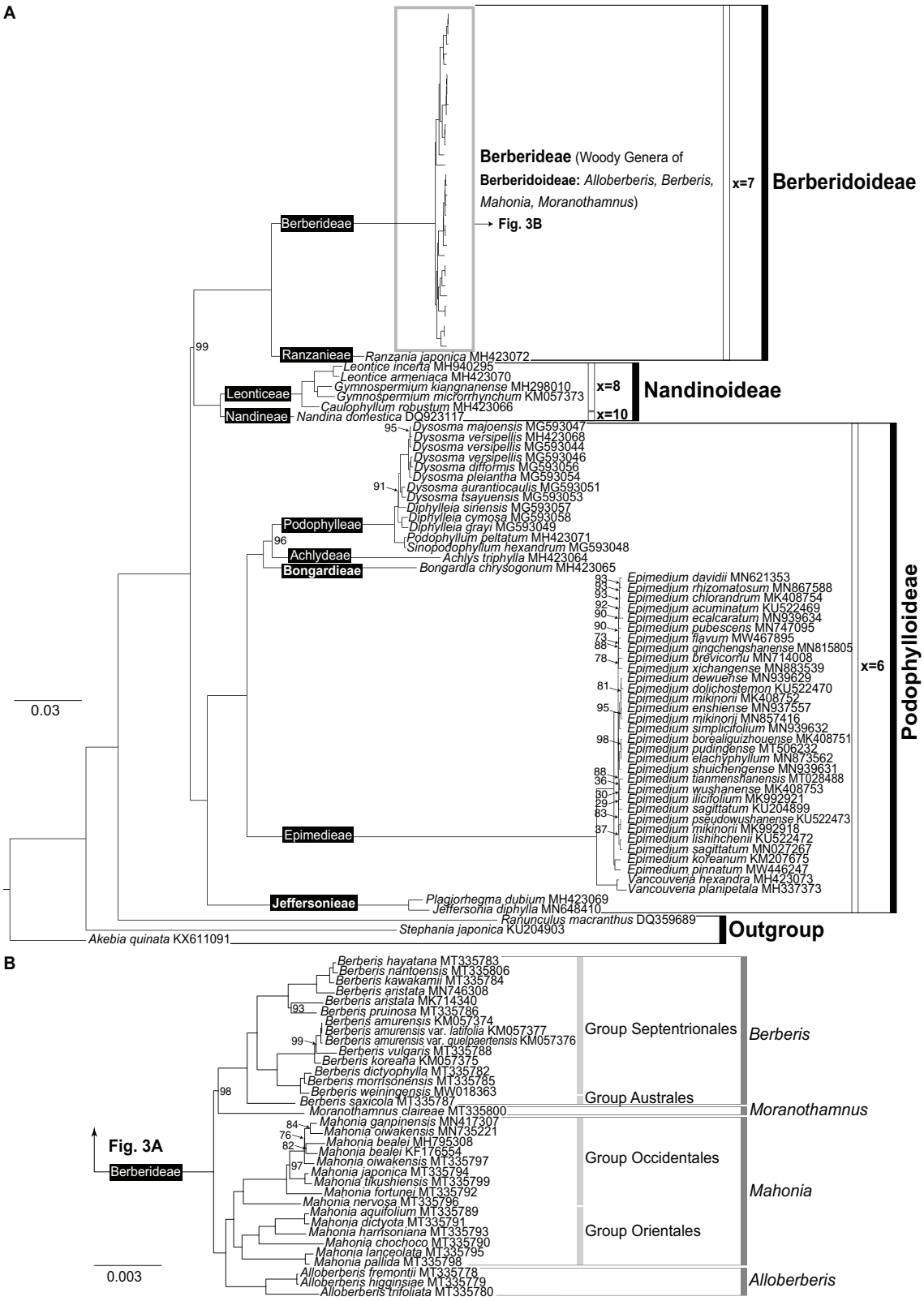


FIGURE 3 | Best-scoring maximum likelihood tree based on the whole plastome sequences reconstructed under IQ-TREE. All branches are fully supported (UFBS: 100) except for those marked by exact numbers (UFBS <100). **(A)** ML phylogram of Berberidaceae labeled with infrafamilial classification proposed in current study; **(B)** phylogenetic relationships of Berberideae with labels following Yu and Chung (2017).

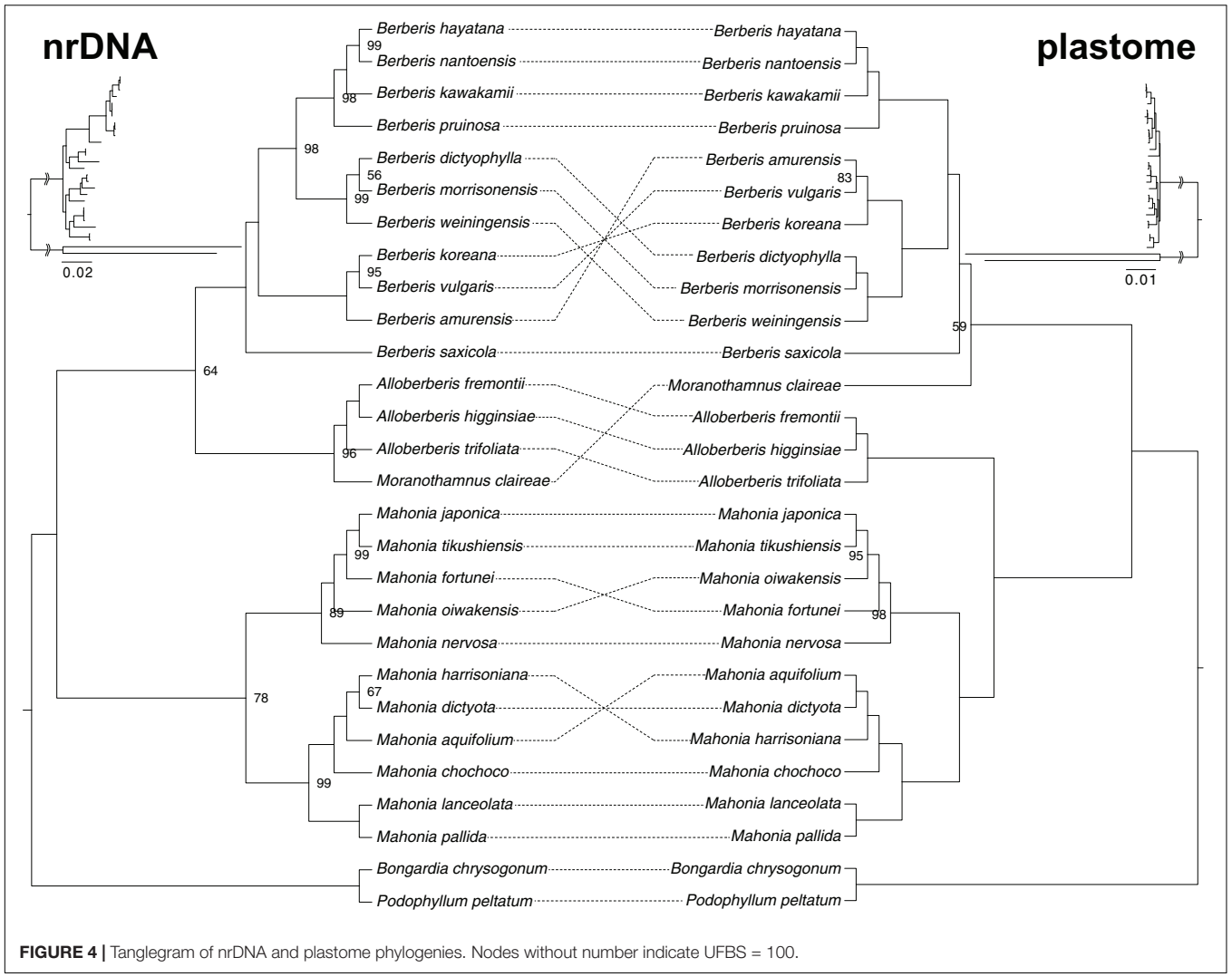


FIGURE 4 | Tanglegram of nrDNA and plastome phylogenies. Nodes without number indicate UFBS = 100.

With IRB excluded and partitions concatenated, alignment of the remaining 93 plastomes is 153,461 bp in length, and contains 36,726 parsimony informative sites (23.93% of the alignment). The substitution models for each partition determined by ModelFinder are listed in **Supplementary Table 6**. Rooted by *A. quinata*, ML phylogeny shows that Ranunculaceae (*Ranunculus macranthus*) is sister to Berberidaceae (**Figure 3A**) with a strong support (UFBS: 100). Within Berberidaceae, three clades corresponding to the three subfamilies Berberidoideae, Nandinoideae, and Podophylloideae were recovered with full support (UFBS: 100).

Within Nandinoideae, *Nandina* (i.e., Nandineae; $x = 10$) is sister to the clade of *Leontice* + *Gymnospermium* + *Caulophyllum* (i.e., Leonticeae; $x = 8$). Within Podophylloideae, five long-branched clades (**Figure 3A**) corresponding to clade *Jeffersonia* + *Plagiorhegma* (i.e., Jeffersonieae *tri. nov.*), clade *Epimedium* + *Vancouveria* (i.e., Epimedieae), *Bongardia* (i.e., Bongardieae *tr. nov.*), *Achlys* (i.e., Achlydeae), and clade *Dysosma* + *Diphylleia* + *Podophyllum* + *Sinopodophyllum* (i.e., Podophylleae) were recovered, with each successive sister to the remaining clades within the subfamily (**Figure 3A**). Within Epimedieae, the monophyletic WNA *Vancouveria* is sister to the monophyletic Eurasian *Epimedium*. Within *Epimedium*, interspecific relationships in general are poorly supported and different relationships have been recovered between the 110-plastome (**Supplementary Figure 1**) and 93-plastome datasets (**Figure 3**); however, in both datasets, *E. pinnatum* (Subgenus *Rhizophyllum*) and *E. koreanum* (Sect. *Macroceras*) form a strongly supported clade sister to the clade of Sect. *Diphyllon* (UFBS: 100). Within Sect. *Diphyllon*, two moderately to strongly supported clades A and B each characterized by slightly different IRB/LSC boundaries were recovered (**Supplementary Figure 9**). Within Podophylleae, *Podophyllum* and *Sinopodophyllum* form a clade sister to *Dysosma* + *Diphylleia*, though *Diphylleia* is paraphyletic with *D. sinensis* sister to *Dysosma* (**Figure 3A**).

Within Berberidoideae (**Figure 3A**), our ML analysis also reveals that *R. japonica* (i.e., Ranzanieae) is sister to Berberideae that is composed of four clades corresponding

to *Alloberberis*, *Berberis*, *Mahonia*, and *Moranothamnus* with full supports, confirming Yu and Chung's (2017) classification. However, while *Alloberberis* was resolved as the sister clade of *Berberis* + *Mahonia* + *Moranothamnus* in Yu and Chung (2017), the genus was placed as the sister group of *Mahonia* with full support in current analysis. Although our sampling of *Berberis* is too limited to test the infrageneric classification of *Berberis* and *Mahonia* (Ahrendt, 1961), the monophyly of Group Septentrionales sister to Group Australes is strongly supported (**Figure 3B**). Within *Mahonia*, the monophyly of the New World Group Occidentales and the predominant Old World Group Orientales are also both fully supported (**Figure 3B**).

Nuclear Ribosomal DNA and Analyses

Table 2 and **Supplementary Table 7** summarizes details of the nrDNA assembly. The average coverage of each species, which was calculated by read mapping, ranges from 876.4× in *M. chochoco* to 3325.9× in *M. japonica* (**Table 2**). The final matrix consists of 7530 aligned base pairs with 855 parsimony informative sites (11.35% of the alignment). The best-fit substitution model for each partition under the best-fit partition scheme was determined by ModelFinder (**Supplementary Table 7**).

Rooted by *Bongardia* and *Podophyllum*, ML analysis of nrDNA using IQ-TREE supports the monophyly of *Alloberberis*, *Berberis*, and *Mahonia* (**Figure 4**), though the support for the monophyly of *Mahonia* is low (UFBS: 78). Within Berberideae, *Mahonia* is sister to the clade composed of *Berberis* + *Alloberberis* + *Moranothamnus*, with the clade *Alloberberis* + *Moranothamnus* sister to *Berberis* with low support (UFBS: 64). As shown in **Figure 4**, relationships in nrDNA tree among the four genera of Berberideae are in conflict with the plastome tree in which *Alloberberis* and *Moranothamnus* are placed sister to *Mahonia* and *Berberis*, respectively, though the support for the latter sister relationship is low (UFBS: 59).

Divergence Time Estimation

The best-fit substitution model and partition scheme as the site model prior for BEAST2 analyses evaluated by ModelFinder were summarized in **Supplementary Table 7**. Using BEAST2, the stem and crown ages of Berberidaceae were estimated to be 91.69 Ma (95% HPD: 103.18–79.93 Ma) and 81.57 Ma (95% HPD: 93.08–70.71 Ma), respectively, falling within the Late Cretaceous (**Table 3** and **Figure 5**). Within Podophylloideae, tribe Jeffersonieae *tr. nov*. diversified from the rest of Podophylloideae at ca. 74.62 Ma (95% HPD: 86.50–62.97 Ma), with the split between *Jeffersonia* and *Plagiorhegma* at ca. 23.15 Ma (95% HPD: 52.72–3.65 Ma). Within the clade of the remaining Podophylloideae, Epimedieae separated from Bongardieae + Achlydeae + Podophylleae at ca. 65.41 Ma (95% HPD: 78.04–53.40 Ma), while Bongardieae split from Achlydeae + Podophylleae at ca. 52.09 Ma (95% HPD: 66.53–36.22 Ma). The split between Achlydeae and Podophylleae was estimated at ca. 43.68 Ma (95% HPD: 57.10–28.36 Ma). The split of Berberidoideae from Nandinoideae was estimated to have occurred at ca. 76.54 Ma (95% HPD: 88.47–64.26 Ma). Within Nandinoideae, Nandineae diverged from the Leonticeae at ca. 48.70 Ma (95% HPD: 73.76–24.75 Ma). Within Berberidoideae, the crown age of Berberidoideae was estimated at ca. 62.24 Ma (95% HPD: 72.09–52.20 Ma). The crown ages of the clades *Alloberberis* + *Mahonia* and *Berberis* + *Moranothamnus* were estimated at ca. 36.23 Ma (95% HPD: 41.36–31.41 Ma) and ca. 32.89 Ma (95% HPD: 43.13–20.59 Ma), respectively. The crown ages of *Alloberberis*, *Berberis*, and *Mahonia* were estimated to be ca. 33.42 Ma (95% HPD: 36.77–31.32 Ma), 20.47 Ma (95% HPD: 33.09–10.08 Ma), and 28.30 Ma (95% HPD: 33.67–23.14 Ma), respectively.

TABLE 3 | Summary of divergence times estimated for genera, tribes, and subfamilies of Berberidaceae by BEAST2.

	Crown age (myr)	Stem age (myr)
Berberidaceae	81.57 (93.08–70.71)	91.69 (103.18–79.93)
Berberidoideae	62.24 (72.09–52.00)	76.54 (88.47–64.26)
Berberideae	38.67 (44.93–32.89)	62.24 (72.09–52.00)
Berberis + *Moranothamnus*	32.89 (43.13–20.59)	38.67 (44.93–32.89)
Moranothamnus	N/A	32.89 (43.13–20.59)
Berberis	20.47 (33.09–10.08)	32.89 (43.13–20.59)
Mahonia + *Alloberberis*	36.23 (41.36–31.41)	38.67 (44.93–32.89)
Alloberberis	33.42 (36.77–31.32)	36.23 (41.36–31.41)
Mahonia	28.30 (33.67–23.14)	36.23 (41.36–31.41)
Ranzanieae (*Ranzania*)	N/A	62.24 (72.09–52.00)
Nandinoideae	48.70 (73.76–24.75)	76.54 (88.47–64.26)
Nandineae (*Nandina*)	N/A	48.70 (73.76–24.75)
Leonticeae	26.66 (40.65–12.10)	48.70 (73.76–24.475)
Caulophyllum	N/A	26.66 (40.65–12.10)
Leontice + *Gymnospermium*	18.21 (30.94–7.55)	26.66 (40.65–12.10)
Leontice	8.04 (17.82–1.27)	18.21 (30.94–7.55)
Gymnospermium	10.70 (21.69–2.43)	18.21 (30.94–7.55)
Nandinoideae + Berberidoideae	76.54 (88.47–64.26)	81.57 (93.08–70.71)
Podophylloideae	74.62 (86.50–62.97)	81.57 (93.08–70.71)
Podophylleae	21.61 (32.90–12.07)	43.68 (57.10–28.36)
Dysosma + *Diphylleia*	16.59 (26.19–9.37)	21.61 (32.90–12.07)
Dysosma	11.47 (17.94–5.75)	14.08 (23.65–7.70)
Podophyllum + *Sinopodophyllum*	5.15 (12.33–0.49)	21.61 (32.90–12.07)
Podophyllum	N/A	5.15 (12.33–0.49)
Sinopodophyllum	N/A	5.15 (12.33–0.49)
Achlydeae (*Achlys*)	N/A	43.68 (57.10–28.36)
Bongardieae (*Bongardia*)	N/A	52.09 (66.53–36.22)
Epimedieae	20.93 (29.58–13.02)	65.41 (78.04–53.41)
Epimedium	13.43 (19.03–7.84)	20.93 (29.58–13.02)
Vancouveria	7.58 (14.72–0.87)	20.93 (29.58–13.02)
Jeffersonieae	23.15 (52.72–3.65)	74.62 (86.50–62.97)
Jeffersonia	N/A	23.15 (52.72–3.65)
Plagiorhegma	N/A	23.15 (52.72–3.65)

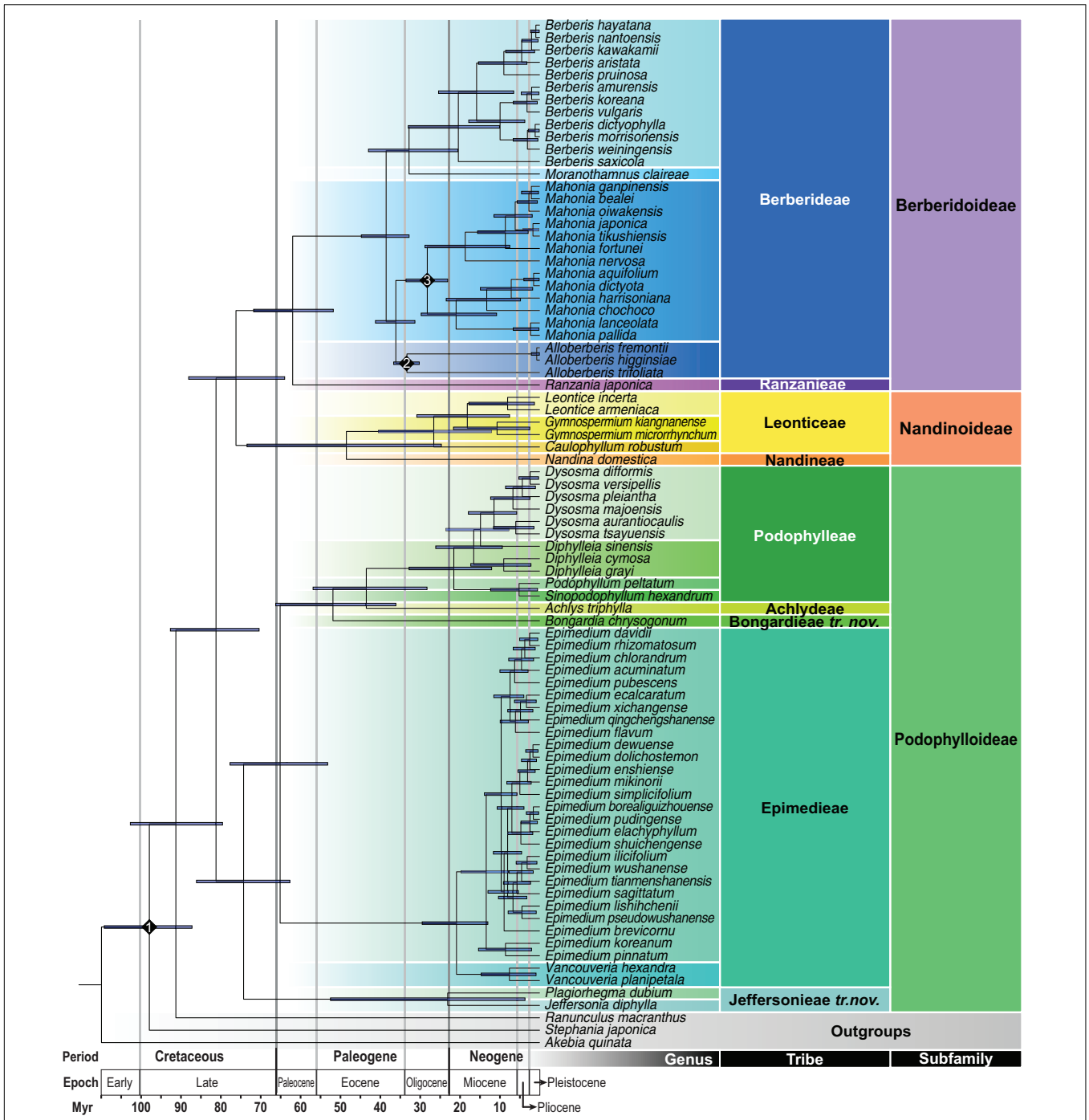


FIGURE 5 | Maximum clade credibility chronogram estimated by BEAST2. The 95% highest posterior density (HPD) date ranges are shown by the node bars. Numbered diamonds are three calibration points.

DISCUSSION

Variation in Berberidoideae Plastome Structure

Despite the functional importance of chloroplasts in photosynthesis and ostensibly the conserved nature of plastid genomes in both structures and contents (Mower and Vickrey, 2018), IR expansions/contractions have been reported across land plants (Goulding et al., 1996; Zhu et al., 2016). In Berberidaceae, early chloroplast restriction site mapping study (Kim and Jansen, 1994) had revealed IR expansion in *Berberis* and *Mahonia* (including *Alloberberis*). Kim and Jansen's (1994)

observation was attested first by the whole plastome sequence of *M. bealei* (Ma et al., 2013) and subsequent phylogenomic analyses (Sun et al., 2018). In current study, all 23 newly assembled plastomes of *Alloberberis*, *Berberis*, and *Mahonia*, as well as the genus *Moranothamnus* that has never been sampled previously, are featured by significant IR expansions (**Supplementary Figure 6**), further corroborating previous studies. However, IR expansion was also reported in *R. japonica* (MG234280) by Wang et al. (2018), contradicting to its conspecific plastome sequence MH423072 (Sun et al., 2018) and early chloroplast restriction site mapping study (Kim and Jansen, 1994). Although the inclusion of MG234280 did not affect phylogenetic relationships of *R. japonica* with the rest of Berberidaceae (**Supplementary Figure 1**), further investigation (e.g., PCR validation) is urgently needed to clarify the SC/IRs junctions in its plastome sequence. Additionally, our analyses also revealed IR expansion and contraction in *E. ecalcaratum* (MN939634) and *E. brevicornu* (MN381716; Zheng et al., 2019), respectively, that have never been reported previously in *Epimedium*. However, in MN803415 (Yao et al., 2020) and MN714008 (Zhang et al., 2020) that are conspecific with MN381716 (Zheng et al., 2019), IR contraction is not detected (**Supplementary Figures 6, 9**). Further study will be needed to clarify the plastome structure in *E. brevicornu* specifically and *Epimedium* in general.

In addition to IR expansion, substantial length variation in *accD* gene featured by insertions and deletions of repeat sequences was revealed in all sampled Berberidoideae plastomes (**Supplementary Figures 7, 8**). AccD encodes the β-carboxyl transferase subunit of acetyl-CoA carboxylase (ACCase), which is a functionally essential multi-subunit enzyme in charge of the biosynthesis of fatty acids in plants (Kode et al., 2005) including non-photosynthetic parasitic plants (e.g., Su et al., 2019). However, pseudogenized *accD* has been reported in *Primula sinensis* (Liu T.J. et al., 2016) and *Vaccinium macrocarpon* (Fajardo et al., 2013). Additionally, *accD* has been lost independently from chloroplast genomes and relocated to the nucleus in gymnosperms, i.e., gnetophytes (Sudianto and Chaw, 2019) and *Sciadopitys verticillata* (Li et al., 2016), and multiple angiosperm species of Acoraceae (Goremykin et al., 2005), Campanulaceae (Hong et al., 2017), Fabaceae (Magee et al., 2010), Geraniaceae (Guisinger et al., 2008), Oleaceae (Lee et al., 2007), and Poales (Harris et al., 2013). Despite the extensive length variation, *accD* genes in Berberidoideae appear to be functional as their reading frames are intact without frameshift and the residual sequences at the 3′ end are highly conserved (**Supplementary Figure 7**). Such length variation characterized by repeat sequences in *accD* has also been reported in the legume species *Medicago truncatula* (Gurdon and Maliga, 2014) and the cupressophytes (Li et al., 2018). Gurdon and Maliga (2014) attributed the intragenic expansion and contraction of *accD* in *M. truncatula* to the presence of repeat sequences that could have triggered replication slippage. Li et al. (2018) also hypothesized that the presence of *accD* repeat sequences could have promoted the acceleration of substitution rate and mediated the rearrangement of plastomes in cupressophytes. Further analyses will be conducted to understand the intriguing *accD* length variation in Berberidoideae.

Ancient Origins of Berberidaceae Genera

Calibrated by fossils of Menispermaceae from the Turonian and *Alloberberis* and *Mahonia* from the Oligocene, the stem age of Berberidaceae was estimated to be 91.69 Ma (95% HPD: 103.18–79.93 Ma), largely congruent with that estimated by Magallón et al. (2015) at 80.28 Ma (95% HPD: 95.84–68.17 Ma), Li et al. (2019) at 87.4 Ma (95% HPD: 98.9–72.9 Ma), and Ramírez-Barahona et al. (2020) at 101.21 Ma (95% HPD: 117.74–87.28 Ma; constrained calibration of a complete set 238 fossils). However, our estimated crown ages of the three subfamilies are much older than those estimated by Sun et al. (2018) [Berberidoideae: 62.24 (95% HPD: 72.09–52 Ma) vs. ca. 16 Ma (95% HPD: 28–6 Ma); Nandinoideae: 48.70 (95% HPD: 73.76–24.75) vs. ca. 24 Ma (95% HPD: 33–13 Ma); Podophylloideae: 74.62 (95% HPD: 86.50–62.97) vs. ca. 32.5 Ma (95% HPD: 36–27 Ma)], in which the divergence times were estimated by constraining the minimum age of the crown group of Berberidaceae at 33.9 Ma. The disparity of age estimates between Sun et al. (2018) and a majority of studies including current one reflects the dubious application of the *Mahonia* fossil to calibrate a deeper node in the former study, resulting in underestimates of ages within the family (Donoghue and Benton, 2007). Indeed, Sun et al. (2018) adopted Magallón et al.'s (2015) calibration strategy that applied the upper Eocene (33.9 Ma) fossil of *Mahonia* as the minimum crown age of Berberidaceae, apparently underestimating the age for the family. Given this, our results (**Table 3** and **Figure 5**) present a more reliable divergence time estimation of the infrafamilial taxa of Berberidaceae than those of Sun et al. (2018).

Within Berberidaceae, our estimated stem ages of genera range from 5.12 Ma (95% HPD: 13.43–0.34 Ma) in *Podophyllum* and *Sinopodophyllum* to 59.74 Ma (95% HPD: 68.56–51.41 Ma) in *Ranzania* (**Table 3** and **Figure 5**). Except for the former two genera that splitted in the early Pliocene, all genera of Berberidaceae were estimated to have originated prior to the early Miocene. While early Pliocene origins of *Podophyllum* and *Sinopodophyllum* are consistent with Liu et al. (2002; 6.52 ± 1.98 Ma) and Wang et al. (2007; 5.8 ± 0.6 Ma), Yu and Chung (2017) has estimated 20.46 Ma (95% HPD: 34.56–2.66 Ma) and 13.78 Ma (95% HPD: 24.47–1.7 Ma) for the stem ages of *Podophyllum* and *Sinopodophyllum*, respectively.

The late Cretaceous origins of the three subfamilies of Berberidaceae estimated in present study are consistent with recent studies of temperate eudicots (e.g., Hypericaceae, Juglandaceae, and Ranunculaceae) in which major lineage diversification had occurred during the Late Cretaceous and Paleocene (Nürk et al., 2015; He et al., 2021; Zhang et al., 2021). Considering the paleoclimate of the Cretaceous, our dating estimation also suggests that the early lineages of Berberidaceae should have adapted to warmer environments, implying niche shifts experienced by extant species (Folk et al., 2020). Notably, the unusually long branch between stem and crown ages of Epimedieae within Podophylloideae also suggests the occurrence of extinction and/or rapid diversification if the sampling bias is ignored (Antonelli and Sanmartín, 2011). Additionally, while

stem ages of most genera within each subfamily were estimated during the Oligocene and Early Miocene, our result suggests the association between the rise of these genera and global climatic deterioration (i.e., temperature cooling and enhanced seasonality) since the Neogene (Smith and Donoghue, 2010). In contrast, the later divergence between *Podophyllum* and the montane *Sinopodophyllum* may be more likely related to the uplift history of the Pan-Himalayan region (Xing and Ree, 2017).

Conflicts Between Plastome and Nuclear Phylogenies

With the inclusion of *Alloberberis* and *Moranothamnus* and expanded sampling of *Berberis* and *Mahonia*, our plastome phylogenomic analyses support the monophyly of Berberideae, its sister relationship with *Ranzania*, and the monophyly of Berberidoideae, Nandinoideae, and Podophylloideae (**Figure 3**), corroborating infrafamilial classification in Berberidaceae (Wang et al., 2009; Yu and Chung, 2017; Sun et al., 2018). However, while previous (Sun et al., 2018) and our current plastome trees both place Berberidoideae sister to Nandinoideae (**Figure 3**), Nandinoideae was resolved as the sister group of Podophylloideae in the combined ML tree of Yu and Chung (2017) and the recently released Kew Tree of Life (KToL) reconstructed using the Hyb-Seq Angiosperms 353 bait set (Johnson et al., 2019; Baker et al., 2021). The conflicting subfamilial relationships of Berberidaceae could have resulted from multiple causes including sampling issues, incomplete lineage sorting (ILS), and hybridization/introgression (Wendel and Doyle, 1998). Given the congruent results between the nuclear trees, i.e., ITS (Yu and Chung, 2017) and the Angiosperms 353 bait set (Baker et al., 2021), the conflicting relationships between plastomes and nuclear datasets observed at the deep level in Berberidaceae seems more likely due to hybridization in the ancient time (Stull et al., 2020).

Within the tribe Berberideae, our phylogenomic analyses (**Figure 3**) reveal four clades corresponding to *Alloberberis*, *Berberis*, *Mahonia*, and *Moranothamnus*, supporting Yu and Chung's (2017) classification. However, relationships of the four genera differ between the plastome and the nrDNA phylogenies (**Figure 4**), as well as Yu and Chung (2017). Specifically, while *Mahonia* and *Alloberberis* are placed in one clade sister to clade *Berberis* + *Moranothamnus* in the plastome tree, in the nrDNA tree *Alloberberis* and *Moranothamnus* formed a strongly supported clade (UFBS: 96) sister to *Berberis*, with *Mahonia* further sister to the clade *Berberis* + *Alloberberis* + *Moranothamnus* (**Figure 4**). Although ILS could have led to this phylogenetic incongruence (Wendel and Doyle, 1998), the conflicting relationships between plastome and nrDNA tree can also be explained by hybridization between *Berberis* and *Mahonia* (García et al., 2017). Under this scenario, *Berberis* and *Mahonia* should be the maternal parents for *Moranothamnus* and *Alloberberis* (**Figure 3**), respectively, given cytoplasmic DNA is known to be maternally inherited in Berberidaceae (Zhang et al., 2003). Coupled with the ancient splits of the four genera (**Figure 5**), *Alloberberis* and *Moranothamnus* could have resulted from ancient reciprocal hybridization (Popelka et al., 2019) between *Berberis* and *Mahonia* preceding subsequent radiations of the two parental genera (García et al., 2017). Additionally, the hybrid origins of *Alloberberis* and *Moranothamnus* could also explain their combined morphology and more restricted geographic distributions relative to *Berberis* and *Mahonia* (Yu and Chung, 2017). Given that *Alloberberis* and *Moranothamnus* are both distributed in western North America, the ancestral ranges of *Berberis* and *Mahonia* are likely also in the New World, as suggested in recent biogeographic study (Chen et al., 2020). Because contemporary intergeneric hybrids between *Berberis* and *Mahonia* (×*Mahoberberis*) rarely occur naturally (Ahrendt, 1961; Rounsaville and Ranney, 2010), the proposition on the hybrid origins of *Alloberberis* and *Moranothamnus* implies a weaker reproductive isolation between the two parental genera in the ancient time.

Within Podophylloideae, although relationships among the five major clades (**Figure 3A**) are largely congruent with previous studies (Wang et al., 2007; Sun et al., 2018), substantial conflicts exist within tribe Podophylleae between current and previous studies. First, in current and four previous studies (Wang et al., 2007; Sun et al., 2018; He et al., 2019; Li and Dong, 2020), *Sinopodophyllum* is placed sister to *Podophyllum*; however, *Sinopodophyllum* was resolved as sister to *Diphylleia* + *Dysosma* + *Podophyllum* in Ye et al. (2018) and *Dysosma* + *Diphylleia* in Yu and Chung (2017). Second, while our plastome tree resolves *Diphylleia* as a paraphyletic grade sister to *Dysosma*, *Dysosma* was resolved sister to *Diphylleia* + *Podophyllum* in Ye et al. (2018), paraphyletic grade sister to *Diphylleia* + *Podophyllum* + *Sinopodophyllum* in He et al. (2019) and Li and Dong (2020), and polyphyletic in Mao et al. (2016). Third, while all three samples species of *Diphylleia* form a clade in Ye et al. (2018), He et al. (2019), and Li and Dong (2020), the genus is paraphyletic in Mao et al. (2016) and current study (**Figure 3A**). One important issue that could have contributed to the conflicting results is the very different strategies utilized to analyze the plastome sequences. In Sun et al. (2018) and Ye et al. (2018), only protein-coding genes (CDS) were analyzed, while He et al. (2019) and Li and Dong (2020) used whole plastomes for phylogenetic reconstruction. Because rates of molecular evolution are in general slower in woody species than the herbaceous members of the same taxonomic group (Smith and Donoghue, 2008; Smith and Beaulieu, 2009), we used the full plastome sequences specifically to increase phylogenetic resolution within Berberidoideae. Another factor that might lead to conflicting relationships is partitioning (Kainer and Lanfear, 2015). While no information regarding partitioning were reported in Ye et al. (2018), He et al. (2019), and Li and Dong (2020), we partitioned the plastome sequences into CDS, RNA regions, introns, and IGS, with CDS further partitioned into three parts by codon positions, to take into account rate variation. Additionally, while the monophyly of *Diphylleia* was supported by a combined tree of cpDNA (*matK* and *rbcL*) and ITS2 (Wang et al., 2007) and ITS (Mao et al., 2014), *matK* and *rbcL* alone did not provide enough phylogenetic information for the monophyly of *Diphylleia* in Wang et al. (2007). Interestingly, Bayesian phylogenetic analysis of *CYP719A*,

a podophyllotoxin biosynthesis gene that could have experienced relaxed purifying selection, showed that both *Diphylleia* and *Dysosma* are not monophyletic (Mao et al., 2016). These conflicting relationships within Podophylleae again could have resulted from ILS and/or hybridization.

To examine whether ILS or hybridization has contributed to conflicting phylogenetic relationships between plastome and nuclear phylogenies, a robust species tree reconstructed from multi-locus genome data (Morales-Briones et al., 2018) such as Angiosperm 353 bait set (Johnson et al., 2019) could provide a promising solution to resolve conflict phylogenetic relationships between plastome and nuclear genes (Shee et al., 2020).

Infrafamilial Classification of Berberidaceae

Based on the robust (**Figure 3**) and dated phylogenomic relationships (**Figure 5**) reconstructed using completed generic sampling of plastome sequences of Berberidaceae, we evaluate different generic concepts outlined in **Supplementary Table 1** using criteria advocated by Backlund and Bremer (1998), Linder et al. (2010), and Heenan and Smissen (2013). Accordingly, our current plastome phylogenomic study (**Figures 3**, **5**) corroborates the classification of four genera within Berberideae (Yu and Chung, 2017). Within Podophylleae, although *Diphylleia* is paraphyletic in our plastome tree (**Figure 3**), the apparent morphological (**Figure 1M**), ecological, phytochemical, anatomical, cytological, and palynological coherence of the genus (Ying et al., 1984; Stearn, 2002) and monophyly as revealed by ITS trees (Wang et al., 2007; Mao et al., 2014) also favor the generic status of this long-recognized genus, though hybridization probably also had occurred in the past. We also support the generic status of the EA *Sinopodophyllum* given its morphological (**Figure 1L**), geographic, and evolutionary distinctness (**Figure 5**) from the ENA *Podophyllum* (Ying, 1979). As a member of the earliest diversified clade (i.e., Jeffersonieae) sister to the rest of Podophylloideae, the generic status of *Plagiorhegma* should also be maintained given its early Miocene split from *Jeffersonia* (**Figure 5**) and morphological (**Figure 1S**) and geographic uniqueness (Hutchinson, 1920).

The maintenance of the generic status of *Alloberberis*, *Mahonia*, *Moranothamnus*, *Plagiorhegma*, and *Sinopodophyllum* that are often synonymized (**Supplementary Table 1**) not only acknowledges their morphological, ecological, and evolutionary distinctness, but also underscores the critical conservation status of these genera. Since the 17th century, *Berberis* has been a major target for eradication around the world because barberry species (and a few species of *Mahonia*) are alternative hosts of rust fungi (Peterson, 2018; Barnes et al., 2020). However, *Alloberberis* (Breckenridge, 1983; Harms, 2007), *Moranothamnus* (Moran, 1982), and a majority of Asian *Mahonia* are highly endangered threatened by habitat destruction and overexploitation (Boufford, 2013) for traditional Chinese medicines (He and Mu, 2015). Subsuming *Alloberberis*, *Mahonia*, and *Moranothamnus* under a broadly defined *Berberis s.l.* would likely further exacerbate their critical conservation status given the stereotypical impression of *Berberis* as agricultural weeds. Additionally, because both *P. dubium* (Lee et al., 2018) and *S. hexandrum* (Liu W. et al., 2016) are also rare and exploited for traditional medicines, recognizing and elevating these two distinct species to the generic rank also confers an effective conservation strategy.

Throughout the taxonomic history of Berberidaceae, several tribes (Janchen, 1949; Terabayashi, 1985b; Loconte, 1993; Takhtajan, 1997; Wu et al., 2003) had been proposed; however, tribal classification has not been implemented under a molecular phylogenetic context. Based on our phylogenomic analyses (**Figures 3**, **5**), we propose to recognize nine clades as tribes within Berberidaceae. We consulted Reveal's (1955–onward) "Indices Nominum Supragenericorum Plantarum Vascularium" for priority of the tribal names. Within Berberidoideae, we follow Terabayashi (1985b) and Wu et al. (2003), recognizing tribes Berberideae (including *Alloberberis*, *Berberis*, *Mahonia*, and *Moranothamnus*) and Ranzanieae (including *Ranzania*). Within Nandinoideae, tribes Leonticeae (including *Caulophyllum*, *Gymnospermium*, and *Leontice*) and Nandineae (including *Nandina*) have long been recognized (**Supplementary Table 1**) and thus are followed here. These two tribes are also characterized by chromosome numbers $x = 8$ and $x = 10$, respectively. Within Podophylloideae, we propose to recognize the five distinct and long-branched clades as tribes (**Figures 3**, **5**). However, while the names Achlydeae, Epimedieae, and Podophylleae are available, the designation Bongardieae (Takhtajan, 1997) was not validly published according to the *Code* (Turland et al., 2018) and the clade *Jeffersonia* + *Plagiorhegma* has never been named. We provide a description for the valid publication of Bongardieae and propose the tribe Jeffersonieae for the latter clade.

Key to Subfamilies, Tribes, and Genera of Berberidaceae

1. Stamens sensitive; pollen exine psilate and imperforate.....2 (Berberidoideae)
1. Stamens not sensitive; pollen exine sculptured and perforate.....6
2. Herbaceous.....Ranzanieae (*Ranzania*)
2. Woody..... 3 (Berberideae)
3. Stem dimorphic..... 4
3. Stem monomorphic..... 5
4. Stem spineless; leaves 3–9-foliolate..... *Alloberberis*
4. Stem almost always spiny; leaves unifoliolate.....*Berberis*
5. Leaves imparipinnate, 5–40-foliolate..... *Mahonia*
5. Leaves uni- to 7-foliolate.....*Moranothamnus*
6. Chromosome base number $x = 8$ or 10..... 7 (Nandinoideae)
6. Chromosome base number $x = 6$..... 10 (Podophylloideae)
7. Woody..... Nandineae (*Nandina*)
7. Herbaceous..... 8 (Leonticeae)
8. Rhizomatous; inflorescence cymose, bracts subulate; flowers calyculate.....*Caulophyllum*
8. Tuberous; inflorescence a raceme or panicle; bracts foliaceous; flowers excalyculate..... 9
9. Leaf solitary, stipulate; seeds exposed by papery pericarp..... *Gymnospermium*

9. Leaves 2–4, sheathing, seeds enclosed in an inflated bladder.....*Leontice*
10. Perianth abscent .....Achlydeae (*Achlys*)
10. Perianth present..... 11
11. Leaves pinnate, with more than six pinnae .....Bongardieae (*Bongardia*)
11. Leaves simple, lobed, or ternately compound..... 12
12. Nectaries absent; aril present..... 13
12. Nectaries present; aril absent .....16 (Podophylleae)
13. Evergreen, petiolules presence, multicellular leaf pubescence present; more than one flowers in an inflorescence.....14 (Epimedieae)
13. Deciduous petiolules absence, multicellular leaf pubescence abscent; one flower in an inflorescence.....15 (Jeffersonieae)
14. Leaves cauline and basal, margins spinose; flowers 2-merous, stamens 4.....*Epimedium*
14. Leaves basal, margins not spinose; flowers 3-merous, stamens 6..... *Vancouveria*
15. Leaves compound; stamens 8..... *Jeffersonia*
15. Leaves simple; stamens 6..... *Plagiorhegma*
16. All leaves with petiole attached to the leaf base..... *Sinopodophyllum*
16. All leaves peltate..... 17
17. Anther dehiscence valvate; ovule anatropous..... *Diphylleia*
17. Anther dehiscence longitudinal; ovule hemitropous..... 18
18. Flowers several in fascicle; stamens 6..... *Dysosma*
18. Flowers solitary; stamens more than 8..... *Podophyllum*

Conspectus of the Infrafamilial Classification of Berberidaceae

Subfamily Berberidoideae Eaton (1836)
- Tribe Berberideae Rchb. (1832)
  - *Alloberberis* C.C.Yu & K.F.Chung, *Berberis* L., *Mahonia* Nutt., *Moranothamnus* C.C.Yu & K.F.Chung
- Tribe Ranzanieae Kumaz. ex Terab. (1985)
  - *Ranzania* T.Ito

Subfamily Nandinoideae Heintze (1927)
- Tribe Leonticeae (Spach) Kosenko (1980)
  - *Caulophyllum* Michx., *Gymnospermium* Spach, *Leontice* L.
- Tribe Nandineae Bernh. (1833)
  - *Nandina* Thunb.

Subfamily Podophylloideae Eaton (1836)
- Tribe Achlydeae Bernh. (1833)
  - *Achlys* DC.
- Tribe Bongardieae Takht. ex C.L.Hsieh, C.C.Yu & K.F.Chung, **tr. nov.**
  - *Bongardia* C.A.Mey.
- Tribe Epimedieae Dumort. (1829)
  - *Epimedium* L., *Vancouveria* C.Morren & Decne.
- Tribe Jeffersonieae C.L.Hsieh, C.C.Yu & K.F.Chung, **tr. nov.**
  - *Jeffersonia* Barton, *Plagiorhegma* Maxim.
- Tribe Podophylleae DC. (1817)
  - *Diphylleia* Michx., *Dysosma* Woodson, *Podophyllum* L., *Sinopodophyllum* T.S.Ying

Tribe **Bongardieae** Takht. ex C.L.Hsieh, C.C.Yu & K.F.Chung, **tr. nov.** – Type: *Bongardia* C.A.Mey.

Bongardieae Takht., Diversity and Classification of Flowering Plants 91. 1997, *num. nud*.

Diagnosis. – Perennial herbs, tuberous. *Tuber* subglobose. *Leaves* glabrous, somewhat fleshy, petiolate, imparipinnate with 7–17 leaflets; leaflets sometimes in whorls of 3 or 4, sessile, obovate to oblong, glaucous-green, usually coarsely toothed from the tip. *Inflorescence* a loose panicle with long scape, 20–60 cm tall. *Flowers* long-stalked; sepals 6, concave, suborbicular or ovate, caducous; petals 6, yellow, oblong–ovate, lanceolate or elliptic-oblong, tips sometimes irregularly crenate. *Stamens* 6. *Ovary* with 5–6 basal ovules, ovoid. *Fruit* a capsule, ovoid, papery, opening from the top by short, acute valves; seeds 1–4, black, pruinose.

Accepted genus. – This tribe contains one genus *Bongardia* C.A.Mey., which is distributed from southern Greece, northern Africa, Middle East to as far east as Pakistan.

Note. – As far as we can track, Bongardieae was first seen in Takhtajan (1997), reiterated in Takhtajan (2009), and adopted by Wu et al. (2003) and Lu and Tang (2020). However, when Takhtajan (1997) published Bongardieae, he did not provide a clear indication of the rank (*Code* Article 37.1), a description/diagnosis (*Code* Article 38.1) in Latin (*Code* Article 39.1), nor a type designation (*Code* Article 40.1). Consequently, Bongardieae Takht. (1997) was not validly published and thus the designation is a *nomen nudum* (Turland et al., 2018).

Tribe **Jeffersonieae** C.L.Hsieh, C.C.Yu & K.F.Chung, **tr. nov.** – Type: *Jeffersonia* Barton

Diagnosis. – Perennial herbs, rhizomatous, deciduous. *Rhizome* short, slender; aerial stems absent. *Leaves* basal; petiole long, slender; leaf blade suborbicular or reniform-orbicular in overall outline, simple or divided into 2 sessile leaflets, palmately veined, margin entire or shallowly lobed. *Flowers* scapose, solitary. *Sepals* 3 or 4, caducous. *Petals* 6 or 8, obovate, pale-purple or white. *Stamens* 6 or 8, antipetalous. *Ovary* with many ovules, placentation marginal. *Fruit* a capsule, dehiscing transversely or longitudinally; seeds numerous.

Accepted genera. – This tribe contains two monotypic genera, *Jeffersonia* Barton and *Plagiorhegma* Maxim., which are disjunctly distributed in eastern North America and East Asia (northeastern China, South Korea, and Russia along Amur River), respectively.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in **Supplementary Table 2**.

AUTHOR CONTRIBUTIONS

K-FC designed the research and provided the research resources. C-CY, Y-LH, and K-FC collected plant materials. Y-LH collected the genomic data. C-LH and Y-LH assembled the plastomes. C-LH analyzed the data and prepared the figures and tables. C-LH, C-CY, and K-FC wrote the manuscript. All authors read and confirmed the manuscript.

FUNDING

This project was supported by the Minister of Science and Technology, Taiwan (MOST 106-2621-B-001-003-MY3).

ACKNOWLEDGMENTS

The authors thank Julian F. Harber, Missouri Botanical Garden Herbarium, Peckerwood Garden, and Rancho Santa Ana Botanic Garden for providing plant materials, the Genomics Core Lab of the Institute of Molecular Biology, Academia Sinica and the High Throughput Genomic Core Lab of Biodiversity Research Center for sequencing, and Bart O'Brien, Takuro Ito, and Mu-Tan Hsieh for permission to use their photographs.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.720171/full#supplementary-material

REFERENCES

Adachi, J., Kosuge, K., Denda, T., and Watanabe, K. (1995). Phylogenetic relationships of the Berberidaceae based on partial sequences of the *gap*A gene. *Plant Syst. Evol.* 9, 351–353. doi: 10.1007/978-3-7091-6612-3_37

Adhikari, B., Milne, R., Pennington, R. T., Särkinen, T., and Pendry, C. A. (2015). Systematics and biogeography of *Berberis* s.l. inferred from nuclear ITS and chloroplast *ndhF* gene sequences. *Taxon* 64, 39–48. doi: 10.12705/641.21

Adhikari, B., Pendry, C. A., and Möller, M. (2014). New chromosome counts of *Berberis* L. (Berberidaceae) suggest that polyploid does not play a significant role in the diversification of the genus in the Nepal Himalaya. *Edinburgh J. Bot.* 71, 297–308. doi: 10.1017/S0960428614000158

Ahrendt, L. W. A. (1961). *Berberis* and *Mahonia*. A taxonomic revision. *Bot. J. Linn. Soc.* 57, 1–410. doi: 10.1111/j.1095-8339.1961.tb00889.x

Airy Shaw, H. K. (1973). *J. C. Willis' A Dictionary of the Flowering Plants and Ferns*, 8th Edn. Cambridge, England: Cambridge University Press.

Anderson, C. L., Bremer, K., and Friis, E. M. (2005). Dating phylogenetically basal eudicots using *rbcL* sequences and multiple fossil reference points. *Am. J. Bot.* 92, 1737–1748. doi: 10.3732/ajb.92.10.1737

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Available Online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Antonelli, A., and Sanmartín, I. (2011). Mass extinction, gradual cooling, or rapid radiation? Reconstructing the spatiotemporal evolution of the ancient angiosperm genus *Hedyosmum* (Chloranthaceae) using empirical and simulated approaches. *Syst. Biol.* 60, 596–615. doi: 10.1093/sysbio/syr062

Backlund, A., and Bremer, K. (1998). To be or not to be—principles of classification and monotypic plant families. *Taxon* 47, 391–400. doi: 10.2307/1223768

Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., et al. (2021). A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Syst. Biol.* 2021:syab035. doi: 10.1093/sysbio/syab035

Barnes, G., Saunders, D. G. O., and Williamson, T. (2020). Banishing barberry: The history of *Berberis vulgaris* prevalence and wheat stem rust incidence across Britain. *Plant Pathol.* 69, 1193–1202. doi: 10.1111/ppa.13231

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

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *Peer J.* 4:e1660. doi: 10.7717/peerj.1660

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Boufford, D. E. (2013). *Mahonia* (Berberidaceae) in Asia: typification, synonymy, and notes. *Mem. N.Y. Bot. Gard.* 108, 251–283.

Breckenridge, F. G. I. (1983). *Berberis swaseyi*—Facing natural extinction? *Texas Native Pl. Soc. News* 1:1.

Brückner, C. (2000). Clarification of the carpel number in Papaverales, Capparales, and Berberidaceae. *Bot. Rev.* 66, 155–307. doi: 10.1007/Bf02858151

Carlquist, S. (1995). Wood anatomy of Berberidaceae: Ecological and phylogenetic considerations. *Aliso* 14, 85–103.

Chen, X.-H., Xiang, K.-L., Lian, L., Peng, H.-W., Erst, A. S., Xiang, X.-G., et al. (2020). Biogeographic diversification of *Mahonia* (Berberidaceae): Implications for the origin and evolution of East Asian subtropical evergreen broadleaved forests. *Mol. Phylogenet. Evol.* 151:e106910. doi: 10.1016/j.ympev.2020.106910

Christenhusz, M. J. M., and Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa* 261, 201–217. doi: 10.11646/phytotaxa.261.3.1

Christenhusz, M. J. M., Fay, M. F., and Byng, J. W. (2018). *The Global Flora: In GLOVAP Nomenclature Part 1*, Special Edn, Vol. 4. Bradford: Plant Gateway Ltd.

Colin, O., Hinsinger, D. D., and Strijk, J. S. (2021). *Mahonia lancasteri* (Berberidaceae), a new species originating from Sichuan (China) described from cultivation. *Phytotaxa* 482, 45–54. doi: 10.11646/phytotaxa.482.1.5

Donoghue, P. C. J., and Benton, M. J. (2007). Rocks and clocks: calibrating the Tree of Life using fossils and molecules. *Trends Ecol. Evol.* 22, 424–431. doi: 10.1016/j.tree.2007.05.005

Doweld, A. B. (2018). New names of fossil Berberidaceae. *Phytotaxa* 351, 72–80. doi: 10.11646/phytotaxa.351.1.6

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Fajardo, D., Senalik, D., Ames, M., Zhu, H. Y., Steffan, S. A., Harbut, R., et al. (2013). Complete plastid genome sequence of *Vaccinium macrocarpon*: structure, gene content, and rearrangements revealed by next generation sequencing. *Tree Genet. Genomes* 9, 489–498. doi: 10.1007/s11295-012-0573-9

Fedde, F. (1901). Versuch einer monographie der gattung *Mahonia*. *Bot. Jahrb. Syst.* 31, 30–133.

Folk, R. A., Siniscalchi, C. M., and Soltis, D. E. (2020). Angiosperms at the edge: Extremity, diversity, and phylogeny. *Plant Cell Environ.* 43, 2871–2893. doi: 10.1111/pce.13887

Fukuda, I. (1967). The biosystematics of *Achlys*. *Taxon* 16, 308–316. doi: 10.2307/1216381

Galasso, G., Conti, F., Peruzzi, L., Ardenghi, N. M. G., Banfi, E., Celesti-Grapow, L., et al. (2018). An updated checklist of the vascular flora alien to Italy. *Plant Biosyst.* 152, 556–592. doi: 10.1080/11263504.2018.1441197

García, N., Folk, R. A., Meerow, A. W., Chamala, S., Gitzendanner, M. A., de Oliveira, R. S., et al. (2017). Deep reticulation and incomplete lineage sorting obscure the diploid phylogeny of rain-lilies and allies (Amaryllidaceae tribe Hippeastreae). *Mol. Phylogenet. Evol.* 111, 231–247. doi: 10.1016/j.ympev.2017.04.003

Gitzendanner, M. A., Soltis, P. S., Yi, T.-S., Li, D.-Z., and Soltis, D. E. (2018). "Plastome phylogenetics: 30 Years of inferences into plant evolution," in *Advances in Botanical Research*, Vol. 85, eds S.-M. Chaw and R. K. Jansen (Cambridge: Academic Press), 293–313. doi: 10.1016/bs.abr.2017.11.016

Goremykin, V. V., Holland, B., Hirsch-Ernst, K. I., and Hellwig, F. H. (2005). Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. *Mol. Biol. Evol.* 22, 1813–1822. doi: 10.1093/molbev/msi173

Goulding, S. E., Olmstead, R. G., Morden, C. W., and Wolfe, K. H. (1996). Ebb and flow of the chloroplast inverted repeat. *Mol. Gen. Genet.* 252, 195–206. doi: 10.1007/Bf02173220

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238

Guisinger, M. M., Kuehl, J. N. V., Boore, J. L., and Jansen, R. K. (2008). Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18424–18429. doi: 10.1073/pnas.0806759105

Gurdon, C., and Maliga, P. (2014). Two distinct plastid genome configurations and unprecedented intraspecies length variation in the *accD* coding region in *Medicago truncatula*. *DNA Res.* 21, 417–427. doi: 10.1093/dnares/dsu007

Hao, D.-C. (2018). *Ranunculales Medicinal Plants: Biodiversity, Chemodiversity and Pharmacotherapy*. London: Academic Press, doi: 10.1016/C2017-0-01185-0

Harms, R. T. (2007). A field study of hybridization between *Berberis swaseyi* and *B. trifoliolata* (Berberidaceae) in Hays County, Texas. *Lundellia* 10, 18–31.

Harris, M. E., Meyer, G., Vandergon, T., and Vandergon, V. O. (2013). Loss of the acetyl-CoA carboxylase (*accD*) gene in Poales. *Plant Mol. Biol. Rep.* 31, 21–31. doi: 10.1007/s11105-012-0461-3

He, J., Lyu, R., Luo, Y., Xiao, J., Xie, L., Wen, J., et al. (2021). A phylotranscriptome study using silica gel-dried leaf tissues produces an updated robust phylogeny of Ranunculaceae. *bioRxiv* 2021:454256. doi: 10.1101/2021.07.29.454256

He, J. M., and Mu, Q. (2015). The medicinal uses of the genus *Mahonia* in traditional Chinese medicine: An ethnopharmacological, phytochemical and pharmacological review. *J. Ethnopharmacol.* 175, 668–683. doi: 10.1016/j.jep.2015.09.013

He, P., Ma, Q., Dong, M., Yang, Z., and Liu, L. (2019). The complete chloroplast genome of *Leontice incerta* and phylogeny of Berberidaceae. *Mitochondrial DNA Part B* 4, 101–102. doi: 10.1080/23802359.2018.1536489

Heenan, P. B., and Smissen, R. D. (2013). Revised circumscription of *Nothofagus* and recognition of the segregate genera *Fuscospora*, *Lophozonia*, and *Trisyngyne* (Nothofagaceae). *Phytotaxa* 146, 1–31. doi: 10.11646/phytotaxa.146.1.1

Heywood, V. H., Brummitt, R. K., Culham, A., and Selders, D. (2007). *Flowering Plant Families of the World*. Ontario: Firefly Books.

Hong, C. P., Park, J., Lee, Y., Lee, M., Park, S. G., Uhm, Y., et al. (2017). *accD* nuclear transfer of *Platycodon grandiflorum* and the plastid of early Campanulaceae. *BMC Genomics* 18:e607. doi: 10.1186/s12864-017-4014-x

Huang, J., Su, T., Lebereton-Anberree, J., Zhang, S. T., and Zhou, Z. K. (2016). The oldest *Mahonia* (Berberidaceae) fossil from East Asia and its biogeographic implications. *J. Plant. Res.* 129, 209–223. doi: 10.1007/s10265-015-0775-y

Huang, Y.-L., Tseng, Y.-H., Chung, K.-F., and Yu, C.-C. (2018). Chromosome numbers of *Berberis* Sect. *Wallichianae* from Taiwan: a new basis for taxonomic and evolutionary implications. *Taiwania* 63, 111–118. doi: 10.6165/tai.2018.63.111

Humphreys, A. M., and Linder, H. P. (2009). Concept versus data in delimitation of plant genera. *Taxon* 58, 1054–1074. doi: 10.1002/tax.584002

Hutchinson, J. (1920). *Jeffersonia* and *Plagiorhegma*. *Bull. Misc. Inform.* 1920, 242–245.

Hutchinson, J. (1973). *The Families of Flowering Plants: Arranged According to a New System Based on Their Probable Phylogeny*, 3rd Edn. Oxford: Oxford University Press.

Janchen, E. (1949). Die systematische Gliederung der Ranunculaceen und Berberidaceen. *Denkschr. Akad. Wiss. Wien, Math.-Naturwiss. Kl.* 108, 1–82.

Jensen, U. (1973). "The interpretation of comparative serological results," in *Chemistry in Botanical Classification: Proceedings of the Twenty Fifth Nobel Symposium*, eds G. Bendz and J. Santesson (New York and London: Academic Press), 217–227.

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome. Biol.* 21:e241. doi: 10.1186/s13059-020-02154-5

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigue, L. R., Cowan, R. S., Devault, A., et al. (2019). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* 68, 594–606. doi: 10.1093/sysbio/syy086

Kainer, D., and Lanfear, R. (2015). The effects of partitioning on phylogenetic inference. *Mol. Biol. Evol.* 32, 1611–1627. doi: 10.1093/molbev/msv026

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/Nmeth.4285

Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kim, Y. D., and Jansen, R. K. (1994). Characterization and phylogenetic distribution of a chloroplast DNA rearrangement in the Berberidaceae. *Plant Syst. Evol.* 193, 107–114. doi: 10.1007/Bf00983544

Kim, Y.-D., and Jansen, R. K. (1995). Phylogenetic implications of chloroplast DNA variation in the Berberidaceae. *Plant Syst. Evol.* 9, 341–349. doi: 10.1007/978-3-7091-6612-3_36

Kim, Y. D., and Jansen, R. K. (1996). Phylogenetic implications of rbcL and ITS sequence variation in the Berberidaceae. *Syst. Bot.* 21, 381–396. doi: 10.2307/2419666

Kim, Y.-D., Kim, S.-H., Kim, C.-H., and Jansen, R. K. (2004a). Phylogeny of Berberidaceae based on sequences of the chloroplast gene ndhF. *Biochem. Syst. Ecol.* 32, 291–301. doi: 10.1016/j.bse.2003.08.002

Kim, Y. D., Kim, S. H., and Landrum, L. R. (2004b). Taxonomic and phytogeographic implications from ITS phylogeny in *Berberis* (Berberidaceae). *J. Plant Res.* 117, 175–182. doi: 10.1007/s10265-004-0145-7

Kode, V., Mudd, E. A., Iamtham, S., and Day, A. (2005). The tobacco plastid *acc*D gene is essential and is required for leaf development. *Plant J.* 44, 237–244. doi: 10.1111/j.1365-313X.2005.02533.x

Kreuzer, M., Howard, C., Adhikari, B., Pendry, C. A., and Hawkins, J. A. (2019). Phylogenomic approaches to DNA Barcoding of herbal medicines: Developing clade-specific diagnostic characters for *Berberis*. *Front. Plant Sci.* 10:586. doi: 10.3389/fpls.2019.00586

Kuroki, Y. (1970). Chromosome study in four species of Berberidaceae. *Mem. Ehime Univ., Sect. 2 Nat. Sci., Ser. B.* 6, 215–221.

Lane, A. K., Augustin, M. M., Ayyampalayam, S., Plant, A., Gleissberg, S., Di Stilio, V. S., et al. (2018). Phylogenomic analysis of Ranunculales resolves branching events across the order. *Bot. J. Linn. Soc.* 187, 157–166. doi: 10.1093/botlinnean/boy015

Lee, H.-L., Jansen, R. K., Chumley, T. W., and Kim, K.-J. (2007). Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Mol. Biol. Evol.* 24, 1161–1180. doi: 10.1093/molbev/msm036

Lee, S.-R., Kim, B.-Y., and Kim, Y.-D. (2018). Genetic diagnosis of a rare myrmecochorous species, *Plagiorhegma dubium* (Berberidaceae): Historical genetic bottlenecks and strong spatial structures among populations. *Ecol. Evol.* 8, 8791–8802. doi: 10.1002/ece3.4362

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., et al. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* 5, 461–470. doi: 10.1038/s41477-019-0421-0

Li, J., Gao, L., Chen, S. S., Tao, K., Su, Y. J., and Wang, T. (2016). Evolution of short inverted repeat in cupressophytes, transfer of *accD* to nucleus in *Sciadopitys verticillata* and phylogenetic position of Sciadopityaceae. *Sci. Rep.* 6:e20934. doi: 10.1038/srep20934

Li, J., Su, Y. J., and Wang, T. (2018). The repeat sequences and elevated substitution rates of the chloroplast *accD* gene in Cupressophytes. *Front. Plant Sci.* 9:533. doi: 10.3389/fpls.2018.00533

Li, R., and Dong, M. (2020). The complete plastid genome of *Jeffersonia diphylla* and its phylogenetic position inference. *Mitochondrial DNA Part B* 5, 77–78. doi: 10.1080/23802359.2019.1696250

Linder, H. P., Baeza, M., Barker, N. P., Galley, C., Humphreys, A. M., Lloyd, K. M., et al. (2010). A generic classification of the Danthonioideae (Poaceae). *Ann. MO Bot. Gard.* 97, 306–364. doi: 10.3417/2009006

Liu, J.-Q., Chen, Z.-D., and Lu, A.-M. (2002). Molecular evidence for the sister relationship of the eastern Asia-North American intercontinental species pair in the *Podophyllum* group (Berberidaceae). *Bot. Bull. Acad. Sinica* 43, 147–154.

Liu, T.-J., Zhang, C.-Y., Yan, H.-F., Zhang, L., Ge, X.-J., and Hao, G. (2016). Complete plastid genome sequence of *Primula sinensis* (Primulaceae): structure comparison, sequence variation and evidence for *accD* transfer to nucleus. *Peer J.* 4:e2101. doi: 10.7717/peerj.2101

Liu, W., Wang, J., Yin, D. X., Yang, M., Wang, P., Han, Q. S., et al. (2016). Genetic diversity and structure of the threatened species *Sinopodophyllum*

hexandrum (Royle) Ying. *Genet. Mol. Res.* 15:15028130. doi: 10.4238/gmr.15028130

Loconte, H. (1993). "Berberidaceae," in *The Families and Genera of Vascular Plants, vol. 2, Flowering plants, Dicotyledons, Magnoliid, Hamamelid and Caryophylliid Families*, eds K. Kubitzki, J. G. Rohwer, and V. Bittrich (Berlin: Springer-Verlag), 147–152. doi: 10.1007/978-3-662-02899-5_14

Loconte, H., and Blackwell, W. H. (1985). Intrageneric taxonomy of *Caulophyllum* (Berberidaceae). *Rhodora* 87, 463–469.

Loconte, H., Campbell, L. M., and Stevenson, D. W. (1995). Ordinal and familial relationships of Ranunculid genera. *Plant Syst. Evol.* 9, 99–118. doi: 10.1007/978-3-7091-6612-3_10

Loconte, H., and Estes, J. R. (1989). Phylogenetic systematics of Berberidaceae and Ranunculales (Magnoliidae). *Syst. Bot.* 14, 565–579.

Lowe, T. M., and Chan, P. P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* 44, W54–W57. doi: 10.1093/nar/gkw413

Lu, A., and Tang, Y. (2020). *The Origin and Evolution of Primitive Angiosperms*. Beijing: Science Press.

Ma, J., Yang, B. X., Zhu, W., Sun, L. L., Tian, J. K., and Wang, X. M. (2013). The complete chloroplast genome sequence of *Mahonia bealei* (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms. *Gene* 528, 120–131. doi: 10.1016/j.gene.2013.07.037

MacGinitie, H. D. (1953). Fossil plants of Florissant beds, Colorado. *Publ. Carnegie Inst. Washington* 599, 1–198.

Magallón, S., Gomez-Acevedo, S., Sanchez-Reyes, L. L., and Hernandez-Hernandez, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* 207, 437–453. doi: 10.1111/nph.13264

Magee, A. M., Aspinall, S., Rice, D. W., Cusack, B. P., Semon, M., Perry, A. S., et al. (2010). Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* 20, 1700–1710. doi: 10.1101/gr.111955.110

Mao, Y.-R., Zhang, Y. H., Nakamura, K., Guan, B.-C., and Qiu, Y.-X. (2014). Developing DNA barcodes for species identification in Podophylloideae (Berberidaceae). *J. Syst. Evol.* 52, 487–499. doi: 10.1111/jse.12076

Mao, Y. R., Zhang, Y. H., Xu, C., and Qiu, Y. X. (2016). Comparative transcriptome resources of two *Dysosma* species (Berberidaceae) and molecular evolution of the CYP719A gene in Podophylloideae. *Mol. Ecol. Resour.* 16, 228–241. doi: 10.1111/1755-0998.12415

Meacham, C. A. (1980). Phylogeny of the Berberidaceae with an evaluation of classifications. *Syst. Bot.* 5, 149–172.

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2011). *The CIPRES science gateway: a community resource for phylogenetic analyses," in Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery*, Vol. 41. New York, NY: ACM, doi: 10.1145/2016741.2016785

Minh, B. Q., Nguyen, M. A. T., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30, 1188–1195. doi: 10.1093/molbev/mst024

Morales-Briones, D. F., Liston, A., and Tank, D. C. (2018). Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytol.* 218, 1668–1684. doi: 10.1111/nph.15099

Moran, R. (1982). Berberis claireae, a new species from Baja California; and why not *Mahonia*. *Phytologia* 52, 221–226.

Mower, J. P., and Vickrey, T. L. (2018). "Structural diversity among plastid genomes of land plants," in *Advances in Botanical Research*, Vol. 85, eds S.-M. Chaw and R. K. Jansen (Amsterdam: Elsevier), 263–292. doi: 10.1016/bs.abr.2017.11.013

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nürk, N. M., Uribe-Convers, S., Gehrke, B., Tank, D. C., and Blattner, F. R. (2015). Oligocene niche shift, Miocene diversification—cold tolerance and accelerated speciation rates in the St. John's Worts (*Hypericum*, Hypericaceae). *BMC Evol. Biol.* 15:80. doi: 10.1186/s12862-015-0359-4

Peng, Y., CHen, S.-B., Liu, Y., Chen, S.-L., and Xiao, P.-G. (2006). A pharmacophylogenetic study of the Berberidaceae (s.l.). *Acta Phytotax. Sin.* 44, 241–257. doi: 10.1360/aps040149

Peterson, P. D. (2018). The barberry eradication program in Minnesota for stem rust control: A case study. *Annu. Rev. Phytopathol.* 56, 203–223. doi: 10.1146/annurev-phyto-080417-050133

Popelka, O., Sochor, M., and Duchoslav, M. (2019). Reciprocal hybridization between diploid *Ficaria calthifolia* and tetraploid *Ficaria verna* subsp. *verna*: evidence from experimental crossing, genome size and molecular markers. *Bot. J. Linn. Soc.* 189, 293–310. doi: 10.1093/botlinnean/boy085

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Ramírez-Barahona, S., Sauquet, H., and Magallón, S. (2020). The delayed and geographically heterogeneous diversification of flowering plant families. *Nat. Ecol. Evol.* 4, 1232–1238. doi: 10.1038/s41559-020-1241-3

Reveal, J. L. (1955–onward). *Indices Nominum Supragenericorum Plantarum Vascularium*. Available Online at: http://www.plantsystematics.org/reveal/pbio/fam/allspgnames.html

Rounsaville, T. J., and Ranney, T. G. (2010). Ploidy levels and genome sizes of *Berberis* L. and *Mahonia* Nutt. species, hybrids, and cultivars. *HortScience* 45, 1029–1033. doi: 10.21273/HORTSCI.45.7.1029

Sastri, R. L. N. (1969). Floral morphology, embryology, and relationships of the Berberidaceae. *Austral. J. Bot.* 17, 69–79. doi: 10.1071/BT9690069

Shaw, J. M. H. (2002). "The genus *Podophyllum*," in *The Genus Epimedium and other Herbaceous Berberidaceae including the Genus Podophyllum*, ed. W. T. Stearn (Kew: The Royal Botanic Gardens), 239–314.

Shee, Z. Q., Frodin, D. G., Cámara-Leret, R., and Pokorny, L. (2020). Reconstructing the complex evolutionary history of the Papuasian *Schefflera* radiation through herbariomics. *Front. Plant Sci.* 11:258. doi: 10.3389/fpls.2020.00258

Smith, S. A., and Beaulieu, J. M. (2009). Life history influences rates of climatic niche evolution in flowering plants. *Proc. R. Soc. Lond. Ser. B. Biol. Sci.* 276, 4345–4352. doi: 10.1098/rspb.2009.1176

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

Smith, S. A., and Donoghue, M. J. (2010). Combining historical biogeography with niche modeling in the *Caprifolium* clade of *Lonicera* (Caprifoliaceae, Dipsacales). *Syst. Biol.* 59, 322–341. doi: 10.1093/sysbio/syq011

Stearn, W. T. (1938). *Epimedium* and *Vancouveria* (Berberidaceae), a monograph. *J. Linn. Soc. Bot.* 51, 409–535. doi: 10.1111/j.1095-8339.1937.tb01914.x

Stearn, W. T. (2002). *The Genus Epimedium and other Herbaceous Berberidaceae including the Genus Podophyllum*. Kew: The Royal Botanic Gardens.

Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A., and Smith, S. A. (2020). Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. *Am. J. Bot.* 107, 790–805. doi: 10.1002/ajb2.1468

Su, H.-J., Barkman, T. J., Hao, W. L., Jones, S. S., Naumann, J., Skippington, E., et al. (2019). Novel genetic code and record-setting AT-richness in the highly reduced plastid genome of the holoparasitic plant *Balanophora*. *Proc. Natl. Acad. Sci. U.S.A.* 116, 934–943. doi: 10.1073/pnas.1816822116

Sudianto, E., and Chaw, S. M. (2019). Two independent plastid *accD* transfers to the nuclear genome of *Gnetum* and other insights on acetyl-CoA carboxylase evolution in Gymnosperms. *Genome Biol. Evol.* 11, 1691–1705. doi: 10.1093/gbe/evz059

Sun, G., Dilcher, D. L., Wang, H.-S., and Chen, Z.-D. (2011). A eudicot from the early cretaceous of China. *Nature* 471, 625–628. doi: 10.1038/nature09811

Sun, Y., Moore, M. J., Landis, J. B., Lin, N., Chen, L., Deng, T., et al. (2018). Plastome phylogenomics of the early-diverging eudicot family Berberidaceae. *Mol. Phylogenet. Evol.* 128, 203–211. doi: 10.1016/j.ympev.2018.07.021

Sun, Y. X., Moore, M. J., Zhang, S. J., Soltis, P. S., Soltis, D. E., Zhao, T. T., et al. (2016). Phylogenomic and structural analyses of 18 complete plastomes across nearly all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution. *Mol. Phylogenet. Evol.* 96, 93–101. doi: 10.1016/j.ympev.2015.12.006

Takhtajan, A. (1997). *Diversity and Classification of Flowering Plants*. New York, NY: Columbia University Press.

Takhtajan, A. (2009). *Flowering Plants*, 2nd Edn. New York, NY: Springer, doi: 10.1007/978-1-4020-9609-9

Terabayashi, S. (1985c). Seedling morphology of the Berberidaceae. *Acta Phytotax. Geobot.* 38, 63–74. doi: 10.18942/bunruichiri.KJ00002992235

Terabayashi, S. (1985a). The comparative floral anatomy and systematics of the Berberidaceae I. *Morphol. Mem. Fac. Sci. Kyoto Univ. Ser. Biol.* 10, 73–90.

Terabayashi, S. (1985b). The comparative floral anatomy and systematics of the Berberidaceae II. Systematic consideration. *Acta Phytotax. Geobot.* 36, 1–13. doi: 10.18942/bunruichiri.KJ00001078521

Thorne, R. F. (1992). Classification and geography of the flowering plants. *Bot. Rev.* 58, 225–327. doi: 10.1007/Bf02858611

Thorne, R. F. (2000). The classification and geography of the flowering plants: Dicotyledons of the Class Angiospermae (*Subclasses Magnoliidae*, Ranunculidae, Caryophyllidae, Dilleniidae, Rosidae, Asteridae, and Lamiidae). *Bot. Rev.* 66, 441–647. doi: 10.1007/BF02869011

Thorne, R. F., and Reveal, J. L. (2007). An updated classification of the class Magnoliopsida ("Angiospermae"). *Bot. Rev.* 73, 67–181.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391

Tonti-Filippini, J., Nevill, P. G., Dixon, K., and Small, I. (2017). What can we do with 1000 plastid genomes? *Plant J.* 90, 808–818. doi: 10.1111/tpj.13491

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., et al. (2018). *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017*. Glashütten: Koeltz Botanical Books, doi: 10.12705/Code.2018

Wang, L.-J., Gao, M.-D., Sheng, M.-Y., and Yin, J. (2020). Cluster analysis of karyotype similarity coefficients in *Epimedium* (Berberidaceae): insights in the systematics and evolution. *Phytokeys* 161, 11–26. doi: 10.3897/phytokeys.161.51046

Wang, M. L., Chen, Y., Hina, F., Ohi-Toma, T., and Li, P. (2018). The complete chloroplast genome of *Ranzania japonica*, an endangered species native to Japan. *Conserv. Genet. Resour.* 10, 671–674. doi: 10.1007/s12686-017-0898-7

Wang, W., Chen, Z.-D., Liu, Y., Li, R.-Q., and Li, J.-H. (2007). Phylogenetic and biogeographic diversification of Berberidaceae in the northern hemisphere. *Syst. Bot.* 32, 731–742. doi: 10.1043/06-16.1

Wang, W., Dilcher, D. L., Sun, G., Wang, H.-S., and Chen, Z.-D. (2016). Accelerated evolution of early angiosperms: Evidence from ranunculalean phylogeny by integrating living and fossil data. *J. Syst. Evol.* 54, 336–341. doi: 10.1111/jse.12090

Wang, W., Lu, A.-M., Ren, Y., Endress, M. E., and Chen, Z.-D. (2009). Phylogeny and classification of Ranunculales: Evidence from four molecular loci and morphological data. *Perspect. Plant Ecol. Evol. Syst.* 11, 81–110. doi: 10.1016/j.ppees.2009.01.001

Wang, W., Ortiz, R. D., Jacques, F. M. B., Xiang, X.-G., Li, H.-L., Lin, L., et al. (2012). Menispermaceae and the diversification of tropical rainforests near the Cretaceous-Paleogene boundary. *New Phytol.* 195, 470–478. doi: 10.1111/j.1469-8137.2012.04158.x

Wen, J., Ickert-Bond, S., Nie, Z.-L., and Li, R. (2010). "Timing and modes of evolution of Eastern Asian-North American biogeographic disjunctions in seed plants," in *Darwin's Heritage Today: Proceedings of the Darwin 200 Beijing International Conference*, eds M. Long, H. Gu, and Z. Zhou (Beijing: Higher Education Press), 252–269.

Wendel, J. F., and Doyle, J. J. (1998). "Phylogenetic incongruence: Window into genome history and molecular evolution," in *Molecular Systematics of Plants II: DNA Sequencing*, eds D. E. Soltis, P. S. Soltis, and J. J. Doyle (Dordrecht: Kluwer Academic), 265–296. doi: 10.1007/978-1-4615-5419-6_10

Wicke, S., and Schneeweiss, G. M. (2015). "Next-generation organellar genomics: Potentials and pitfalls of high-throughput technologies for molecular evolutionary studies and plant systematics," in *Next-Generation Sequencing in Plant Systematics*, eds E. Hörandl and M. Appelhans (Oberreifenberg: Koeltz Botanical Books), 1–42. doi: 10.14630/000002

Wu, Z., Lu, A., Tang, Y., Chen, Z., and Li, D. (2003). *The Families and Genera of Angiosperms in China: A Comprehensive Analysis*. Beijing: Science Press.

Xing, Y.-W., and Ree, R. H. (2017). Uplift-driven diversification in the Hengduan Mountains, a temperate biodiversity hotspot. *Proc. Natl. Acad. Sci. U.S.A.* 114, E3444–E3451. doi: 10.1073/pnas.1616063114

Yao, Y., Liu, X., Yang, Q., Luo, Y., Zhang, C., Xu, C., et al. (2020). The complete chloroplast genome of *Epimedium brevicornu* (Berberidaceae), a traditional Chinese medicinal herb. *Mitochondrial DNA Part B* 5, 887–888. doi: 10.1080/23802359.2020.1718027

Ye, W.-Q., Yap, Z. Y., Li, P., Comes, H. P., and Qiu, Y.-X. (2018). Plastome organization, genome-based phylogeny and evolution of plastid genes in Podophylloideae (Berberidaceae). *Mol. Phylogenet. Evol.* 127, 978–987. doi: 10.1016/j.ympev.2018.07.001

Ying, T.-S. (1979). On *Dysosma* Woodson and *Sinopodophyllum* Ying, gen. nov. of the Berberidaceae. *Acta Phytotax. Sin.* 17, 15–23.

Ying, T.-S., Terabayashi, S., and Boufford, D. E. (1984). A monograph of *Diphylleia* (Berberidaceae). *J. Arnold Arbor.* 65, 57–94.

Yu, C.-C. (2018). *Molecular Phylogenetics and Historical Biogeography of Berberis L. (Berberidaceae)*. Ph.D. Dissertation. Taipei: National Taiwan University.

Yu, C.-C., and Chung, K.-F. (2017). Why *Mahonia*? Molecular recircumscription of *Berberis* s.l., with the description of two genera, *Alloberberis* and *Moranothamnus*. *Taxon* 66, 1371–1392. doi: 10.12705/666.6

Zhang, M. L., Uhink, C. H., and Kadereit, J. W. (2007). Phylogeny and biogeography of *Epimedium*/*Vancouveria* (Berberidaceae): Western North American-East Asian disjunctions, the origin of European mountain plant taxa, and East Asian species diversity. *Syst. Bot.* 32, 81–92. doi: 10.1600/036364407780360265

Zhang, M. Y., Lu, L., Wortley, A. H., Wang, H., Li, D. Z., and Blackmore, S. (2017). Evolution of angiosperm pollen: 4. Basal eudicots. *Ann. MO. Bot. Gard.* 102, 141–182. doi: 10.3417/2015035

Zhang, Q., Liu, Y., and Sodmergen. (2003). Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant Cell Physiol.* 44, 941–951. doi: 10.1093/pcp/pcg121

Zhang, Q., Ree, R. H., Salamin, N., Xing, Y., and Silvestro, D. (2021). Fossil-informed models reveal a boreotropical origin and divergent evolutionary trajectories in the Walnut Family (Juglandaceae). *Syst. Biol.* 2021:syab065. doi: 10.1093/sysbio/syab030

Zhang, Y., Huang, R., Wu, L., Wang, Y., Jin, T., and Liang, Q. (2020). The complete chloroplast genome of *Epimedium brevicornu* Maxim (Berberidaceae), a traditional Chinese medicine herb. *Mitochondrial DNA Part B* 5, 588–590. doi: 10.1080/23802359.2019.1710593

Zhang, Y. J., Du, L. W., Liu, A., Chen, J. J., Wu, L., Hu, W. M., et al. (2016). The complete chloroplast genome sequences of five *Epimedium* species: Lights into phylogenetic and taxonomic analyses. *Front. Plant Sci.* 7:306. doi: 10.3389/fpls.2016.00306

Zheng, G., Zhang, C., Yang, J., and Xu, X. (2019). Characterization of the complete chloroplast genome of *Epimedium brevicornu* (Berberidaceae). *Mitochondrial DNA Part B* 4, 3681–3682. doi: 10.1080/23802359.2019.1678429

Zhou, Z.-H. (2014). The jehol biota, an early cretaceous terrestrial lagerstätte: new discoveries and implications. *Natl. Sci. Rev.* 1, 543–559. doi: 10.1093/nsr/nwu055

Zhu, A. D., Guo, W. H., Gupta, S., Fan, W. S., and Mower, J. P. (2016). Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209, 1747–1756. doi: 10.1111/nph.13743

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 07 January 2022
doi: 10.3389/fpls.2021.765719

How to Tackle Phylogenetic Discordance in Recent and Rapidly Radiating Groups? Developing a Workflow Using *Loricaria* (Asteraceae) as an Example

Martha Kandziora[1†], Petr Sklenář[1], Filip Kolář[1,2†] and Roswitha Schmickl[1,2†]*

[1] Department of Botany, Faculty of Science, Charles University, Prague, Czechia, [2] Institute of Botany, The Czech Academy of Sciences, Průhonice, Czechia

OPEN ACCESS

Edited by:
Susann Wicke,
Humboldt University of Berlin, Germany

Reviewed by:
Julia Bechteler,
University of Bonn, Germany
Sidonie Bellot,
Royal Botanic Gardens, Kew, United Kingdom

***Correspondence:**
Martha Kandziora
kandziom@natur.cuni.cz

†ORCID:
Martha Kandziora
orcid.org/0000-0002-1197-6207
Filip Kolář
orcid.org/0000-0002-8793-7992
Roswitha Schmickl
orcid.org/0000-0002-0632-5143

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: 27 August 2021
Accepted: 22 November 2021
Published: 07 January 2022

Citation:
Kandziora M, Sklenář P, Kolář F and Schmickl R (2022) How to Tackle Phylogenetic Discordance in Recent and Rapidly Radiating Groups? Developing a Workflow Using Loricaria (Asteraceae) as an Example. Front. Plant Sci. 12:765719. doi: 10.3389/fpls.2021.765719

A major challenge in phylogenetics and -genomics is to resolve young rapidly radiating groups. The fast succession of species increases the probability of incomplete lineage sorting (ILS), and different topologies of the gene trees are expected, leading to gene tree discordance, i.e., not all gene trees represent the species tree. Phylogenetic discordance is common in phylogenomic datasets, and apart from ILS, additional sources include hybridization, whole-genome duplication, and methodological artifacts. Despite a high degree of gene tree discordance, species trees are often well supported and the sources of discordance are not further addressed in phylogenomic studies, which can eventually lead to incorrect phylogenetic hypotheses, especially in rapidly radiating groups. We chose the high-Andean Asteraceae genus *Loricaria* to shed light on the potential sources of phylogenetic discordance and generated a phylogenetic hypothesis. By accounting for paralogy during gene tree inference, we generated a species tree based on hundreds of nuclear loci, using Hyb-Seq, and a plastome phylogeny obtained from off-target reads during target enrichment. We observed a high degree of gene tree discordance, which we found implausible at first sight, because the genus did not show evidence of hybridization in previous studies. We used various phylogenomic analyses (trees and networks) as well as the D-statistics to test for ILS and hybridization, which we developed into a workflow on how to tackle phylogenetic discordance in recent radiations. We found strong evidence for ILS and hybridization within the genus *Loricaria*. Low genetic differentiation was evident between species located in different Andean cordilleras, which could be indicative of substantial introgression between populations, promoted during Pleistocene glaciations, when alpine habitats shifted creating opportunities for secondary contact and hybridization.

Keywords: rapid radiation, hybridization, workflow, incomplete lineage sorting, gene tree discordance, cytonuclear discordance

INTRODUCTION

While rapidly radiating groups are interesting to science due to their potential to understand evolution, adaptation, and the impact of environmental change on biodiversity, they pose one of the biggest challenges in resolving the tree of life (Whitfield and Lockhart, 2007; Song et al., 2012; Escudero et al., 2020; Morales-Briones et al., 2021). Phylogenies of rapid radiations have short internal branches due to the fast succession of species. This rapid accumulation of species can be the result of different and non-exclusive processes, such as geographic isolation, sexual selection or ecological adaptation (Schluter, 2000; Givnish, 2015). The short time between speciation events in rapid radiations increases the probability of incomplete lineage sorting (ILS), i.e., the phenomenon of ancestral polymorphism persisting between successive speciation events (Maddison, 1997). This potentially reduces phylogenetic signal (Townsend, 2007; Whitfield and Lockhart, 2007), as the different topologies of the gene trees expected under ILS lead to gene tree discordance, i.e., not all gene trees represent the species tree. The advent of phylogenomics has not only brought novel methods of generating large datasets, but also new methods of inferring phylogenetic trees and networks. Several available methods to reconstruct the species tree and phylogenetic networks account for ILS (e.g., Than et al., 2008; Vachaspati and Warnow, 2015; Solís-Lemus et al., 2017; Zhang et al., 2017). ILS can be addressed by using multi-species coalescent (MSC) methods for phylogenetic reconstruction, where the different evolutionary histories of loci are considered. Especially in lineages that show a high degree of ILS, species tree estimations are usually more reliable than concatenation (Jiang et al., 2020).

Besides ILS, other processes can lead to phylogenetic discordance, both among gene trees (hereafter referred to as gene tree discordance) and across genomes (among different genomic compartments within a genome; hereafter referred to as cytonuclear discordance). Within plants these processes are mainly hybridization and whole-genome duplication (WGD; Degnan and Rosenberg, 2009). Hybridization frequently occurs under the form of 'introgressive hybridization,' i.e., the introduction of syntenic nucleotide variation from a donor species into the genome of a recipient species, by means of hybridization and backcrossing (Anderson and Hubricht, 1938). Evidence for hybridization from phylogenetic datasets has traditionally been obtained from cytonuclear discordance (i.e., the incongruence between nuclear and plastome trees) and using graph-based networks (e.g., NeighborNet, SuperQ; Bryant and Moulton, 2004; Grunewald et al., 2013). While these approaches remain applicable in the phylogenomics era, they rather depict reticulation; commonly interpreted as evidence for hybridization, however this is non-exclusive (Huson and Bryant, 2006; Degnan, 2018). In contrast, a few methods that account for ILS and simultaneously address hybridization exist which allow testing of hybrid origin, among them the D-statistics (ABBA-BABA statistics; Patterson et al., 2012) and phylogenetic networks (Than et al., 2008; Solís-Lemus et al., 2017). As these model-based approaches are computationally demanding and feasible only for a small number of samples and putative hybridization events (Kamneva et al., 2017; Folk et al., 2018), testing for hybrid origins in phylogenomic datasets remains a challenge.

Hybrids are frequently meiotically stabilized via WGD, but WGD events also occur in the absence of hybridization. After a WGD event, gene copies are subsequently lost (e.g., Xiang et al., 2017). If duplicated non-homologous sequences are not differentiated into their orthologous pairs, the orthology assumption for phylogenetic reconstruction is violated. Alignments that consist of paralogous sequences may lead to biased phylogenetic inference (Fitch, 1970; Gabaldón, 2008; Yang and Smith, 2014) or not (Yan et al., 2021). The best practice to account for paralogy in phylogenetic reconstruction is under debate. Four strategies can be followed: (1) deleting paralogous loci from the analysis (Jones et al., 2019; Larridon et al., 2020); (2) retrieving both ortho- and paralogous copies of the loci without separating them into different alignments and proceeding with a gene duplication-aware species tree method (Zhang et al., 2020a); (3) retrieving both ortho- and paralogous copies without separating them into different alignments and proceeding with an ILS-aware species tree method (Yan et al., 2021); and (4) retrieving all copies, both ortho- and paralogous, and creating orthologous alignments, from which gene trees are inferred, before building the species tree (Gizaw et al., 2021; this study).

Once gene tree discordance is found in a dataset, its sources should be deciphered, as it can lead to wrong estimations of phylogenetic relationships (Huson and Bryant, 2006). Apart from the evolutionary processes mentioned above, methodological artifacts due to missing data, scarce sampling of taxa, and incorrect model specifications can be additional sources of gene tree discordance (Molloy and Warnow, 2018; Nute et al., 2018). Especially when resolving phylogenetic relationships in rapid radiations, the degree of gene tree discordance is expected to be high, as a large number of loci has to be employed, which increases the likelihood of sampling loci that evolved under ILS, hybridization, and WGD (Degnan and Rosenberg, 2009). As such, rapidly radiating groups present a challenge in terms of resolution in phylogenomic datasets, at least when using currently available computational methods (Esselstyn et al., 2017; Reddy et al., 2017). Studies about the sources of genome and gene tree discordance either focus on evolutionary model organisms, where *a priori* knowledge about putative hybridization events is larger (e.g., Meier et al., 2017; Lee-Yaw et al., 2019) or ancient radiations, where the effect of ILS is decreased due to species extinction (e.g., Rosidae, Sun et al., 2015; Amaranthaceae s.l., Morales-Briones et al., 2021). In contrast, phylogenetic discordance in young radiating groups only recently gained attention: *Lachemilla* Focke (Rydb.) (Morales-Briones et al., 2018), *Lomatium* Raf. (Ottenlips et al., 2021), and *Veronica* L. (Thomas et al., 2021). A lack of data sources, such as phylogenies based on Sanger sequence markers, detailed morphological evaluations, and flow cytometric measurements in combination with chromosome counts, make it particularly difficult to disentangle sources of phylogenetic incongruence in young understudied groups.

Rapidly radiating groups can be found in young biodiversity hotspots, such as the high altitude areas of tropical South America

(Madriñán et al., 2013). Comprehensive sampling of lineages that are either large in species number, cover large geographic areas or include many micro-endemics often pose substantial taxonomic and fieldwork challenges. In the past, the use of a few Sanger sequence markers or the plastome often did not provide sufficient resolution at shallow phylogenetic levels. As such, many of these lineages or genera are understudied and remain poorly understood, including several Andean radiations: e.g., *Astragalus* L. (Bagheri et al., 2017); *Diplostephium* Kunth (Vargas et al., 2017); Espeletiinae (Diazgranados and Barber, 2017; Cortés et al., 2018); *Lupinus* L. (Drummond et al., 2012; Contreras-Ortiz et al., 2018); and *Senecio* L. (Kandziora et al., 2016).

The family Asteraceae is one of the youngest and most species-rich families among the angiosperms and accounts for a large diversity within tropical alpine ecosystems (Sklenář et al., 2011; Panero and Crozier, 2016). WGD events and hybridization are common for many members of the Asteraceae (Smissen et al., 2011; Galbany-Casals et al., 2014; Barker et al., 2016; Huang et al., 2016; Zhang et al., 2021). For this study, we chose the high-Andean genus *Loricaria* Wedd. from the tribe Gnaphalieae as a representative of a young radiating group. The genus comprises 19 species and has an estimated stem age of 6 million years (Ma; crown age of 4 Ma) according to Nie et al. (2016). The genus occurs above 3500m in the tropical Andes from Bolivia to Colombia. During Pleistocene glacial cycles, the tropical alpine ecosystem shifted downwards in the cold and drier periods (Van der Hammen, 1985; Hooghiemstra and Van der Hammen, 2004; Flantua et al., 2019) and species changed their ranges, met, and potentially hybridized. Interestingly, there was no evidence for hybridization in *Loricaria* based on amplified fragment length polymorphism (AFLP) data, and polyploids have not been detected to date (Kolář et al., 2016).

In this study, we addressed phylogenetic relationships among *Loricaria* using Hyb-Seq (Weitemier et al., 2014). We encountered a high degree of discordance, both among gene trees and across genomes, which seems to be common for Asteraceae (Siniscalchi et al., 2021). We then aimed to disentangle ILS, hybridization, and WGD as possible sources of the immense gene tree discordance, and we established a workflow for phylogenetic inference of young radiating groups that accounts for these sources of discordance. We additionally accounted for a possible impact of missing data on gene tree discordance. Our workflow is especially useful for non-model groups, for which often only limited knowledge exists about hybridization events and polyploidy.

MATERIALS AND METHODS

Taxonomic Focus

Loricaria is a genus restricted to the high elevation habitats of the tropical Andes. The genus comprises small dioecious shrubs with scale-like leaves, which is a morphological convergence to certain gymnosperm genera (Cuatrecasas, 1954). Currently, 19 species are accepted, 17 as a result of the synopsis of the genus by Cuatrecasas (1954), and two species described by Dillon and Sagastegui Alva (1986), and Hind (2004). Morphological investigations of herbarium specimens plus signatures of potential cryptic speciation within *L. thuyoides* (Lam.) Sch. Bip. (Kolář et al., 2016) suggested that there is a substantial degree of taxonomic uncertainty and that there potentially exist more species than are described to date.

The genus has been divided into three different sections (**Table 1**), primarily based on the position of the flower heads, which is axillary in sect. *Thyopsis* and sect. *Graveoleum* and terminal in sect. *Terminalia*. Section *Graveoleum* is differentiated from the other two sections by a glandulose-pilose ovary and glandulose-pubescent leaves. *Loricaria graveolens* (Sch. Bip.) Wedd. is the only member of sect. *Graveoleum*. Distribution information and section assignation are taken from Cuatrecasas (1954), Dillon and Sagastegui Alva (1986), and Hind (2004).

The genus belongs to the tribe Gnaphalieae, which has its species richness concentrated in the southern hemisphere (Nie et al., 2016). Hybridization has been inferred for this tribe (Galbany-Casals et al., 2014; Barker et al., 2016), thus phylogenetic discordance can be expected. Further, within the tribe Gnaphalieae the most recent common ancestor (mrca) of the FLAG-clade, which includes *Loricaria* [defined in Galbany-Casals et al. (2010), the acronym stands for the species-rich genera within this clade: *Filago* L., *Leontopodium* R. Br. ex Cass., *Antennaria* Gaertn., and *Gamochaeta* Wedd.], likely underwent a hybridization plus WGD event (Smissen et al., 2011). Further, the Gnaphalieae experienced a WGD event about 10 Ma ago (Huang et al., 2016; Zhang et al., 2020b). These WGD events are expected to add phylogenetic discordance.

Sampling and DNA Sequencing

Sampling was carried out over a period of 12 years (2006–2018) by one of the authors. Leaves were dried on silica for DNA extraction. Herbarium specimens are deposited in PRC, and duplicates stored in QCA, QCNE, and AAU. Further, we got additional material from AAU, B, MA, and BONN. We sampled 15 out of the 19 accepted *Loricaria* species and, in addition, four new morphological groups representing potentially new species (nine individuals; **Supplementary Table 1**).

For Hyb-Seq, we included between 1 and 13 samples per species to test for their monophyly, with a stronger focus on sect. *Thyopsis*, as it includes more widespread species than the other sections. The outgroup was complemented with sequences from Mandel et al. (2019), who used the same probe set for target enrichment (see below). In total, 13 species from seven genera of the Gnaphalieae were sampled as outgroup taxa. Overall, we sampled 63 individuals for this study, including 13 outgroup samples.

DNA extraction, genomic library preparation, and bait hybridization followed Gizaw et al. (2021). We used the Compositae1061 probe set (Mandel et al., 2014), implemented in the myBaits Expert Compositae1061 target capture kit (Arbor Biosciences, Ann Arbor, MI, United States). Enriched libraries were mixed with unenriched libraries in the ratio 2: 1 (run2), 1.5: 1 (run3), and 1: 1 (run5), respectively. Samples were sequenced on different sequencers, either an Illumina (San Diego, CA, United States) NextSeq at the Genomics Core Facility of CEITEC (Brno, Czechia) or an Illumina NovaSeq at IAB

TABLE 1 | Species and characteristics of *Loricaria*.

Species	Section	Clade	Capitulum position	Distribution
L. graveolens (Sch. Bip.) Wedd.	Graveoleum	Graveolens	Axilliary	Peru (Cuatrecasas, 1954)
L. ollgaardii M.O. Dillon & Sagast.	Thyopsis	Unknown	Terminal	Ecuador (Dillon and Sagastegui Alva, 1986)
L. complanata (Sch. Bip.) Wedd.	Thyopsis	Axilliary	Axilliary	Ecuador, Colombia (Cuatrecasas, 1954)
L. thuyoides (Lam.) Sch. Bip.	Thyopsis	Axilliary	Axilliary	Peru, Ecuador, Colombia (Cuatrecasas, 1954)
L. scolopendra (Hook.) Kuntze	Thyopsis	Axilliary	Axilliary	Ecuador (Cuatrecasas, 1954)
L. pauciflora Cuatrec.	Thyopsis	Axilliary	Axilliary	Ecuador (Cuatrecasas, 1954)
L. azuayensis Cuatrec.	Thyopsis	Axilliary	Axilliary	Ecuador (Cuatrecasas, 1954)
L. cinerea D. J. N. Hind	Thyopsis	unknown	Axilliary and terminal	Ecuador (Hind, 2004)
L. lagunillensis Cuatrec.	Thyopsis	unknown	Axilliary	Colombia (Cuatrecasas, 1954)
***L. leptothamna* (Mattf.) Cuatr.**	**Thyopsis**	**(Terminal)**	**Terminal**	Peru (Cuatrecasas, 1954)
L. puracensis Cuatrec.	Terminalia	Terminal	Terminal	Colombia (Cuatrecasas, 1954)
L. lucida Cuatrec.	Terminalia	Unknown	Terminal	Peru (Cuatrecasas, 1954)
L. ferruginea (Ruiz & Pav.) Wedd.	Terminalia	Terminal	Terminal	Peru, Ecuador (Cuatrecasas, 1954)
L. lycopodinae Cuatrec.	Terminalia	Terminal	Terminal	Peru (Cuatrecasas, 1954)
L. antisanensis Cuatrec.	Terminalia	Terminal	Terminal	Ecuador (Cuatrecasas, 1954)
L. ilinissae (Benth.) Cuatrec.	Terminalia	Terminal	Terminal	Ecuador (Cuatrecasas, 1954)
L. macbridei Cuatrec.	Terminalia	Unknown	Terminal	Peru (Cuatrecasas, 1954)
***L. colombiana* Cuatrec.**	**Terminalia**	**Terminal**	**Axilliary**	Colombia (Cuatrecasas, 1954)
***L. unduaviensis* Cuatrec.**	**Thyopsis**	**Terminal**	**(Axilliary) and terminal**	Bolivia (Cuatrecasas, 1954)

In bold are highlighted where the species' sectional assignment does not match their placement in clades and/or position of the capitula.

(Olomouc, Czechia); in all cases 150 base pairs (bp) paired-end reads were obtained. Raw reads are available under NCBI SRA BioProject PRJNA777419.

Data Analysis Workflow

We developed a data analysis workflow that implements data filtering, paralog detection and utilization for phylogenetic reconstruction, and investigation of ILS and hybridization to robustly infer the phylogeny of a young radiating group (**Figure 1**). Most scripts used herein are part of HybPhyloMaker (Fér and Schmickl, 2018; scripts are available at https://github.com/tomas-fer/HybPhyloMaker), which we indicate as "HPM" followed by the number of the respective script. Customizing the reference, paralog detection, and orthologous alignment building was performed using ParalogWizard[1] (scripts are available at https://github.com/rufimov/ParalogWizard), to which we refer to as "PW" followed by the number of the respective script. In the case of running scripts outside of these two bioinformatic pipelines we refer to the scripts directly in the respective methodological section, and if steps need to be done by the user manually, we denote this as "manual." All steps and scripts are also summarized in the **Supplementary Table 3**.

Nuclear Read Assembly, Paralog Identification and Locus Alignment

The first part of the workflow will assemble nuclear reads, identify paralogs and align loci (**Figure 1A**). Raw reads were trimmed to remove adapters and low quality bases using Trimmomatic v.0.39 (Bolger et al., 2014) and duplicates were removed using BBMap v.38.42 (Bushnell, 2014) following the settings implemented in HPM 1: Low quality bases were considered to have a base quality encoding below 33 (phred33) and were coded as N. Additionally, we removed low quality bases at the beginning and end of the read if below Q20, and if bases in a sliding window of 5 bp were below the threshold Q20, the read was cut and the bases removed. Finally, we deleted reads shorter than 36 bp. Reads were assembled into contigs *de novo* using Compositae1061 exons as target file for initial read fishing during the assembly step (distribute_reads_to_targets_bwa.py and spades_runner.py from HybPiper v.1.3.1 [Johnson et al., 2016] implemented in PW 1a and b [see text footnote 1]). The minimum coverage to call a single nucleotide polymorphism (SNP) was set to 2. Subsequently, we customized the target file with sequences from our *Loricaria* reads (PW 2b) and repeated the mapping with this "Loricaria-optimized Compositae1061" target file (PW 1a and b), consisting of the best matching, longest exonic sequences from different *Loricaria* samples (see text footnote 1). Outgroup taxa were specified in the "blocklist" to exclude them in order to generate a target file containing only *Loricaria* sequences.

We followed (see text footnote 1; option 4 from the introduction) to detect paralogous loci and use both paralogs and orthologs for phylogenetic reconstruction. The approach is summarized in Gizaw et al. (2021). In brief, to assess paralogy, pairwise sequence divergence between the exonic contigs of each locus was estimated, which resulted in two main clusters of divergence, the first denoting allelic variation and the second paralogy. Similar to Gizaw et al. (2021), we chose the mean of the second cluster ± the standard deviation as the divergence threshold for considering exonic copies to be paralogous, using

[1] Ufimov, R., Gorospe, J. M., Fér, T., Kandziora, M., Salomon, L., van Loo, M., et al. (in prep). Utilizing paralogs for phylogenetic reconstruction has the potential to increase support and reduce gene tree discordance in target enrichment data. In preparation for Molecular Ecology Resources.

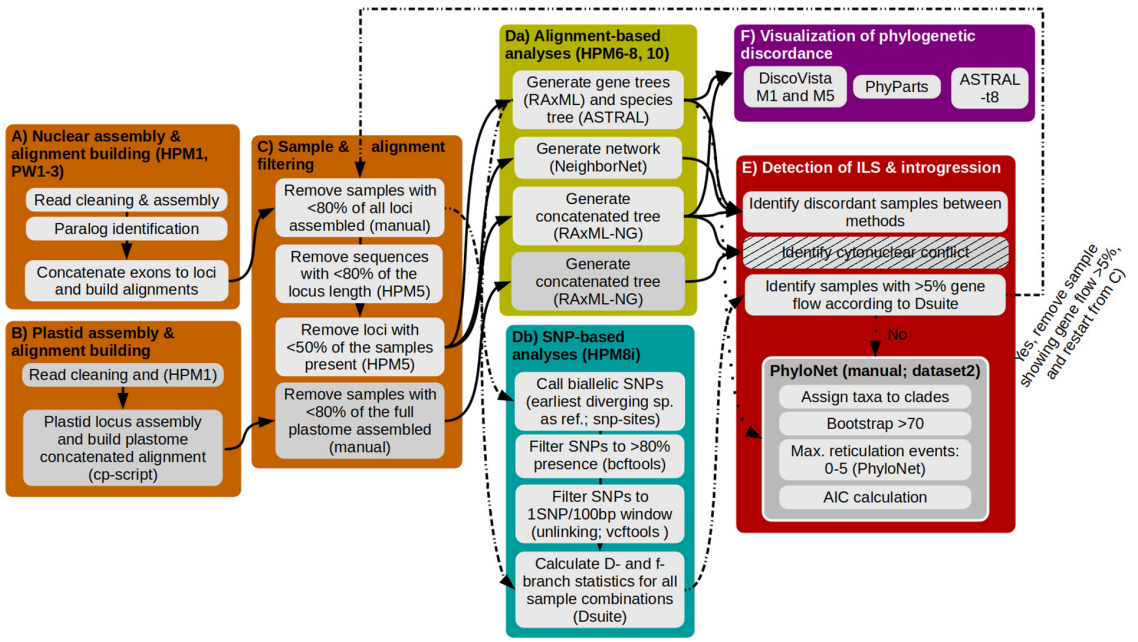


FIGURE 1 | Illustration of the different steps employed for the discovery of samples and clades which show introgression to other clades. Each colored box represents a major analysis step, enumerated from **(A–E)**, analysis scripts used are indicated in the respective boxes in parentheses. **(A)** Assembly of nuclear reads, identification of paralogs, and alignment building. **(B)** Assembly of plastid reads and alignment building. **(C)** Filtration of alignments to exclude samples with few assembled exons and sequences that are too short. **(Da)** Calculation of gene and species trees. **(Db)** Identification and filtration of single nucleotide polymorphisms and analysis of gene flow. **(E)** Identification of ILS and introgression. At first, the pipeline follows the solid arrows which results in dataset 1. This dataset 1 is used to follow the dashed-dotted lines in an iterative approach to remove samples that show gene flow, finally resulting in dataset 2. Using dataset 2 and following the dotted line allows to identify if all hybridigenous samples have been detected. **(F)** Visualization of phylogenetic discordance between phylogenies is done for dataset 1 and dataset 2, respectively.

PW 2a. For our dataset, sequence divergence between 7.96 and 19.43% are considered to represent paralogous copies. Exonic alignments for each orthologous and paralogous copy were built using MAFFT v.7.029 (Katoh et al., 2005) and the exons concatenated to loci using PW 3.

Plastome Read Assembly

The second part of the workflow assembles the plastome and builds a concatenated alignment (**Figure 1B**). Plastome sequence data were obtained as a by-catch as the result of our adapted lab protocol that adds a proportion of unenriched libraries to the enriched libraries. Reads were trimmed for quality using Trimmomatic v.0.39 (Bolger et al., 2014) and duplicates removed using BBMap v.38.42 (Bushnell, 2014) as implemented in HPM 1. Detailed settings about read filtering are provided in the paragraph before. The remaining reads were mapped to a user-provided reference (*Leontopodium*; GenBank accession number NC027835) using BWA v.0.7.15 (Li and Durbin, 2009), implemented in a script available at https://github.com/tomas-fer/scripts/blob/master/cpDNA_mappingMETA.sh. Before the read mapping, one of the two inverted repeats of the plastome reference was removed using Geneious v.2020.1.2[2]. We then called the consensus sequence using kindel v.0.1.4 (Constantinides and Robertson, 2017) with a minimum read depth of 2 and with a 0.51 threshold for consensus variant calling; regions of the plastome without mapped reads were coded as N. The alignment was built using MAFFT.

Sample and Alignment Filtering

The third part of the worklow filters samples and alignments with too much missing data (**Figure 1C**). As missing data have a substantial impact on correct species tree estimation, especially under a high degree of ILS (Nute et al., 2018), we tested different subsampling strategies toward optimally incorporating poorly assembled samples into the nuclear and plastome phylogenies, respectively (hereafter referred to as low quality samples): not excluding them, excluding samples with less than 50% assembled loci, and with less than 80% assembled loci. Based on the number of loci recovered per sample that is reported in the table 'MissingDataOverview.txt' (created by HPM 5), we deleted samples manually from the analyses folder before continuing. We aimed at an optimal tradeoff between the number of assembled loci per sample and the number of samples removed from the dataset due to low quality. As such, the nuclear and plastome datasets include slightly different sets of samples.

We employed a second filtering step, this time for the nuclear alignments only. For each locus, we excluded sequences missing more than 50% data for the locus, and we removed loci for which less than 80% of all samples were represented (HPM 5).

[2]www.geneious.com

Alignment- and Single Nucleotide Polymorphism-Based Analyses and Identification of Incomplete Lineage Sorting and Introgression

The fourth part of the workflow consists of alignment- and SNP-based analyses (**Figure 1D**). Based on the filtered alignments from the step above (hereafter referred to as dataset 1), we inferred phylogenetic hypotheses using two methods, the MSC method (nuclear dataset only) and concatenation (**Figure 1Da**). For the MSC method, gene trees were estimated using RAxML v.8.4.2 (Stamatakis, 2014) with the general-time reversible (GTR) substitution model with a gamma distributed rate variation among sites "GTRGAMMA" and 500 bootstrap replicates (HPM 6a). Based on these gene trees, we generated an ASTRAL species tree using ASTRAL III v.5.6.1 (Zhang et al., 2017; HPM 8a). For the species tree calculation, we initially tested the effect of collapsing poorly supported gene tree nodes (HPM 10). This showed no effect in our dataset, but we recommend to test that during analysis. Additionally, all loci were concatenated into a locus-partitioned supermatrix (HPM 8f) and a phylogeny was inferred using RAxML-NG v.8 (Kozlov et al., 2019; run manual), hereafter referred to as concatenated tree. For the plastome, the same concatenation approach, but without partitioning was utilized.

In a next step, the datasets were evaluated for signatures of ILS and hybridization (**Figure 1E**). The first round of evaluation employed commonly used approaches, which do not provide full evidence of ILS and/or hybridization. Incongruent placement of samples between the following phylogenetic comparisons are commonly treated as an indication of ILS and/or hybridization: First, based on the nuclear phylogenetic reconstructions we identified incongruent placements between the ASTRAL species tree and the concatenated tree. Second, based on a comparison between the plastome and the nuclear (ASTRAL) phylogeny we inferred cytonuclear discordance. In addition, a distance-based network was generated using NeighborNet (Bryant and Moulton, 2004) available in SplitsTree v.4.16.2 (Huson, 1998; HPM 8 h). Admixed samples were determined to be those forming mixed groups (i.e., samples grouping with different samples in the NeighborNet compared to well-supported clades in the phylogenetic results) or showing a misplacement in the network (i.e., isolated samples).

The second round of evaluation was based on full-evidence approaches that simultaneously account for ILS and introgression: Dsuite (HPM 8i; **Figure 1Db**) and PhyloNet (run manual; **Figure 1E**). Using Dsuite v.0.4r38 (Patterson et al., 2012; Malinsky et al., 2021; HPM 8i), we calculated the D-statistics (also called ABBA-BABA statistics) for all trios (with a fixed outgroup) of species in the dataset and f-branch statistic to evaluate the amount of introgression. The D-statistics estimates the frequency of "ABBA" and "BABA" patterns in a four-taxon phylogeny {[(Sample1,Sample2)Sample3]Outgroup}, whereas the SNP "A" denotes the ancestral SNP and "B" the derived. Under ILS, both patterns are equally likely, whereas an excess of one pattern indicates introgression. To perform the analyses, we called SNPs for the ingroup based on *L. graveolens* (sample LR_017; sister to the remaining species in the phylogeny) as reference using snp-sites v.2.3.3 (Page et al., 2016). Before SNP calling we concatenated all loci irrespective of missing data. We retained only biallelic SNPs, and removed SNPs with more than 20% missing data using bcftools v.1.7 (Li, 2011), leading overall to more SNPs than if we would have used the alignments after the missing data filtering step. To only retain unlinked SNPs, we then filtered the biallelic SNPs for one SNP per 100 bp window using vcftools v.0.1.17 (Danecek et al., 2011), resulting in about 1500 SNPs in total from initial 16,000 SNPs. To run the Dsuite analysis, the ASTRAL species tree was used as input. Dsuite allows to assign samples to species to account for intra-specific variation, but we decided to map samples to samples, as most species were not retrieved as monophyletic, and we did not know if this was due to a hybridigenous origin of these species or of certain samples. We will use the term 'hybridigenous origin' throughout the manuscript in cases where we are not able to distinguish between introgression and hybridization. Samples that showed more than 5% introgression to samples from different clades based on the f-branch statistic (introgression cut-off according to Malinsky et al., 2021) were removed and the species tree recalculated. We employed an iterative approach by rerunning the analysis until no major events of introgression (>5%) could be detected, as we found that Dsuite continued detecting samples with signatures of introgression to other samples after removing the first set of samples showing introgression. Removal of all introgressed samples resulted in a reduced dataset (hereafter referred to as dataset 2).

To test if the Dsuite analyses detected all introgressed samples and to reveal if the evolutionary structure of *Loricaria* is tree-like, we applied evolutionary network analyses implemented in PhyloNet v.3.8.2 (Than et al., 2008) using the computationally least demanding maximum pseudo-likelihood (MPL) approach. PhyloNet returns the degree of gene flow, denoted as γ-parameter, of the two parents to the hybrid. We used all gene trees after removing samples that showed introgression according to the Dsuite analyses (dataset 2), and only considered nodes with a bootstrap support greater than 70% (nodes were collapsed prior to analysis using HPM 10). We allowed between zero to five hybridization events. Samples were mapped to supported clades; without this data simplification, none of the analyses finished within 2 weeks calculation time employing 14 CPUs. We then performed a model selection using the Akaike information criterion (AIC); the best supported network is the one with the lowest AIC value. The number of parameters for the AIC calculation equals the number of branches plus the number of allowed reticulation events and number of gene trees used to estimate the likelihood.

Visualization of Gene Tree Discordance Across Different Datasets

As the sixth part of the workflow, gene tree discordance is measured between phylogenies using different approaches (**Figure 1F**). First, we used PhyParts based on all gene trees and a node-support threshold of 70% (Smith et al., 2015; HPM 11). PhyParts calculates the number of concordant as well as discordant nodes between gene trees in comparison to the species tree. Second, we used the quartet-based method provided in ASTRAL III (the -t 8 option; run manual) to

identify the percentage of alternative quartets. For the quartet-based method, an equal proportion of quartets indicates a high degree of ILS (Sayyari and Mirarab, 2018). Third, in contrast to the above mentioned methods, to differentiate between highly and moderately conflicting nodes that conflict between gene trees and the species tree and, hence, to provide a more detailed understanding of the gene tree discordance we used DiscoVista (Sayyari et al., 2018; run manual). DiscoVista also permits to visualize differences in gene tree discordance between our different sample and alignment filtering approaches. The discordance analysis on gene trees (method 1 [M1] in **Figure 1F**) and the relative frequency analysis of alternative topologies (method 5 [M5] in **Figure 1F**) were based on a support threshold value of 70% and a maximum of 20% missing samples in the clade.

RESULTS

Assembly, Alignment, and Alignment Filtering

As missing data have a substantial impact on correct species tree estimation, especially under a high degree of ILS (Nute et al., 2018), we chose a relatively stringent tradeoff between the number of assembled loci per sample and the number of samples removed from the dataset due to low quality. We excluded samples with a recovery of less than 80% across all loci from both the nuclear and plastome datasets, which resulted in the removal of 26 and 28 samples for the nuclear and plastome dataset, respectively. In total, we used three outgroup samples and 34 samples from the ingroup for the nuclear dataset, reducing the dataset from 63 to 37 samples. For the plastome dataset, we used 10 outgroup samples and 25 samples from the ingroup, totaling 35 samples. Alignments were built for each dataset separately, and after sample filtering resulted in 13 species plus our additional four new morphological groups for the nuclear alignments, and 13 species plus additionally three of the new morphological groups of *Loricaria* for the plastome alignments.

After trimming and de-duplicating the raw reads, between 0.9 and 14 million reads per sample were retained. The nuclear dataset 1, from which the initial ASTRAL species tree was reconstructed, consisted of 973–1150 loci per sample. To avoid data loss by removing paralogous loci, we differentiated between orthologous and paralogous copies and built separate alignments from those, which increased the number of loci for phylogenetic reconstruction by about 25%. We did not find a higher degree of paralogy in particular species or clades (**Supplementary Table 2**).

Without employing our stringent sample filtering by removing samples with less than 80% of the exons recovered, the support for clades within *Loricaria* was low [<0.95 local posterior probability (LPP); data not shown]. After removing those samples (HPM 5), the number of loci used during alignment building increased, as these low-quality samples did not longer have a dominating effect on the removal of loci that had less than 80% of all samples present (**Supplementary Figure 1A**). Further, gene tree discordance decreased (**Supplementary Figure 1A**), and the support for the major clades (those discussed later) increased, from only four nodes within *Loricaria* having a support greater than 0.95 LPP to eight such nodes (**Figure 2A**; not counting supported nodes within species-complexes).

The number of mapped plastid reads ranged from 5027 to 199,001, with an average proportion of missing data of 3.3% (min-max: 0.06–17.6%). The length of the concatenated plastome

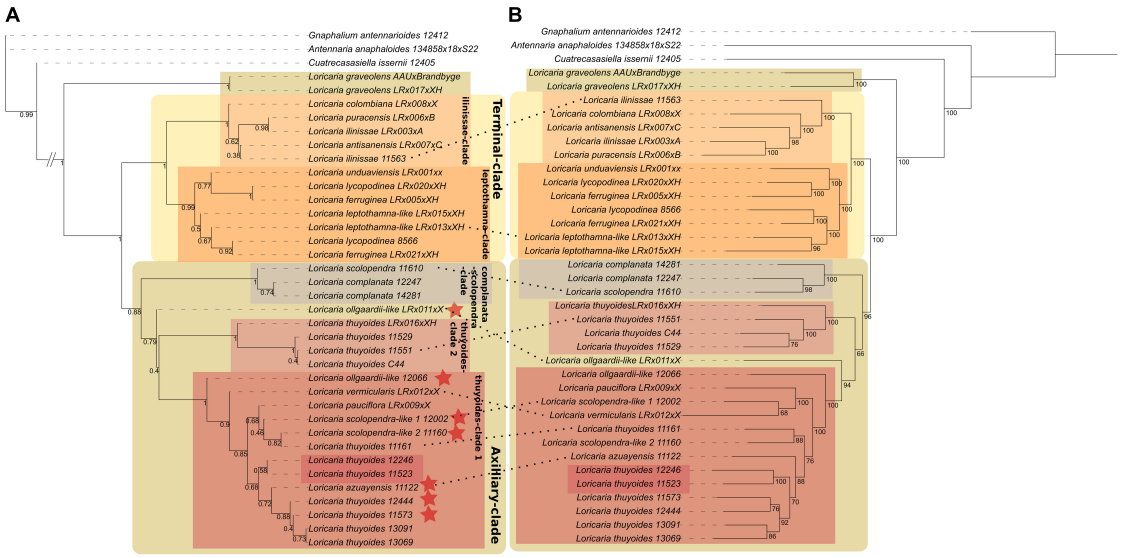


FIGURE 2 | Nuclear species tree using ASTRAL III **(A)** in comparison to the concatenated tree calculated using RAxML-NG (**B**; dataset 1). Values at nodes in **(A)** represent local posterior probabilities and in **(B)** bootstrap values. Samples that differ in their phylogenetic placement are indicated by a dotted line. Red stars indicate samples that got removed from dataset 2. Labeled boxes indicate major clades discussed in the main text. Red unlabeled boxes indicate *L. thuyoides* samples with a different placement in the plastome tree (see **Supplementary Figure 4**).

dataset after removing samples with more than 20% missing data was 261,701 bp.

Phylogenetic Analyses and Testing for Signatures of Incomplete Lineage Sorting

The monophyly of *Loricaria* was strongly supported based on the nuclear data, both in the ASTRAL species tree (1 LPP) and the concatenated tree (100% bootstrap support [BS]; **Figure 2**; dataset 1). The earliest diverging taxon to the outgroup, *L. graveolens*, the only representative of the sect. *Graveoleum,* was the only species with support for monophyly (ASTRAL: 1 LPP; concatenation: 100% BS). Based on the ASTRAL and concatenated tree the two main clades retrieved grouped species of high morphological similarity (for dataset 1: **Figure 2**; for dataset 2: **Supplementary Figure 6**), and we hereafter refer to these main clades as the Terminal- (1 LPP; 100% BS for dataset 1 and 2) and Axilliary-clade (0.88 LPP/0.96 LPP for dataset 1/dataset 2 and 96%/76% BS, respectively). The phylogenetic placement of most samples followed the sectional classification of the genus, with the exception of *L. unduaviensis* Cuatrec. and one new morphological group, *L. "leptothamna-like,"* which were placed with samples from the sect. *Terminalia*. *Loricaria unduaviensis* and *L. leptothamna* were placed in the sect. *Thyopsis* by Cuatrecasas (1954), irrespective of the position of their capitulum, which is terminal, not axilliary. Similarly, *Loricaria colombiana* has axilliary capitulas, but is part of the Terminal-clade and assigned to sect. *Terminalia*. The samples identified as *L. "ollgaardii-like,"* have high morphological similarity to *L. ollgaardii* except for the position of their capitulas, which is axilliary in our samples but terminal in the original species description. Accordingly, our samples were found in the Axilliary-clade. In addition to the *L. "ollgaardii-like"* samples, the Axilliary-clade consisted of two subclades, the scolopendra-complanata-clade (1 LPP; 100% BS for dataset 1 and 2) and the thuyoides-complex, which comprised samples from *L. pauciflora* Cuatrec., *L. azuayensis* Cuatrec., and *L. thuyoides* as well as three of the four new morphological groups. Within the thuyoides-complex, samples of the species *L. thuyoides* were found to be non-monophyletic: samples of *L. pauciflora* and *L. azuayensis* were nested among *L. thuyoides* samples. Further, four *L. thuyoides* samples formed a sister clade relationship to the remaining samples of the Axilliary-clade (1 LPP; 100% BS for dataset 1 and 2), denoted as thuyoides-clade2. The Terminal-clade comprised the ilinissae-clade (1 LPP; 100% BS for dataset 1 and 2), including *L. ilinissae* (Benth.) Cuatrec., *L. puracensis Cuatrec., L. antisanensis* Cuatrec., and *L. colombiana* Cuatrec., and the leptothamna-clade (dataset 1: 0.99 LPP; 100% BS; dataset 2: 0.97 LPP; 100% BS), composed of *L. "leptothamna-like," L. lycopodinae* Cuatrec., *L. ferruginea* (Ruiz & Pav.) Wedd., and *L. unduaviensis*.

We recovered the same clades for the concatenated phylogeny as we did based on the ASTRAL species tree, with only a few samples showing different placements within these clades (**Figure 2**). Analysis of gene tree discordance using PhyParts showed that each node is highly discordant, with only a small proportion of gene trees supporting the species tree (**Supplementary Figure 2**). When investigating the frequency of the different topologies retrieved among the gene trees using DiscoVista (method 5), for the node leading to the genus, one topology was most frequent (>50%, **Figure 3A**). After removing samples that showed gene flow (dataset 2), the split separating the early diverging species *L. graveolens* from the remaining clades was found in more than 50% of the topologies (**Figure 3B**). The remaining nodes showed almost equal frequencies for all three topologies (highest frequency of maximally 40%; **Figure 3A**). The percentage of alternative quartets according to ASTRAL gave similar results, mostly having similar frequencies for all three quartets (**Supplementary Figure 3**). Only the node indicating genus monophyly and the species *L. graveolens* were well supported with more than 50% of the quartets showing the same topology. Clades supported by more than 40% of the quartets were the ilinissae-clade, the scolopendra-complatana-clade, and the thuyoides-clade2, as well as the mrca of the Axilliary- and Terminal-clade (**Supplementary Figure 3**).

Testing for Signatures of Introgression

Cytonuclear discordance was strong in our dataset (**Supplementary Figure 4**). The plastome tree was reconstructed using a slightly different selection of samples than for the nuclear trees, due to our missing data filter approach. While most clades recovered in the nuclear dataset were present in the plastome tree, there were major differences. First, the genus was non-monophyletic: *Belloa schultzii* (Wedd.) Cabrera formed a clade with the earliest diverging species *L. graveolens* (100% BS). Second, the only sample available from the scolopendra-complanata-clade (14281) for the plastome dataset plus two samples of *L. thuyoides* (11523, 12246) formed a clade together with the samples from the ilinissae-clade (100% BS), while other members of the thuyoides-clade1 did not (**Supplementary Figure 4**). Additionally, the ilinissae-clade is part of the Axilliary-clade in the plastome tree (100% BS).

According to the distance-based network, we identified a different set of misplaced samples compared to the ones identified in the plastome tree as well as those that showed signs of ILS: Samples from two of the new morphological groups, i.e., *L. "ollgaardii-like"* samples (12066, LR_011) and the *L. "scolopendra-like 1"* sample (12002; **Supplementary Figure 5**) grouped with different samples in the network than in the phylogenetic reconstructions.

Calling SNPs from our Hyb-Seq data resulted in an initial set of 16,000 SNPs before filtering according to our minimum threshold for SNP presence across samples and accounting for linkage. Approximately 50% of the SNPs were removed due to their presence in less than 80% of the samples, and restricting the SNP dataset to unlinked SNPs resulted in a further reduction to final 1515 SNPs. Based on the Dsuite analyses, we identified seven samples showing introgression greater than 5% (**Table 2**). All these samples belonged to the thuyoides-clade1: three out of the four new morphological groups (*L. "ollgaardii-like"*: 12066, LR_011; *L. "scolopendra-like 1"*: 12002; *L. "scolopendra-like 2"*: 11160), two samples of *L. thuyoides* (12444, 11573), and *L. azuayensis* (11122).

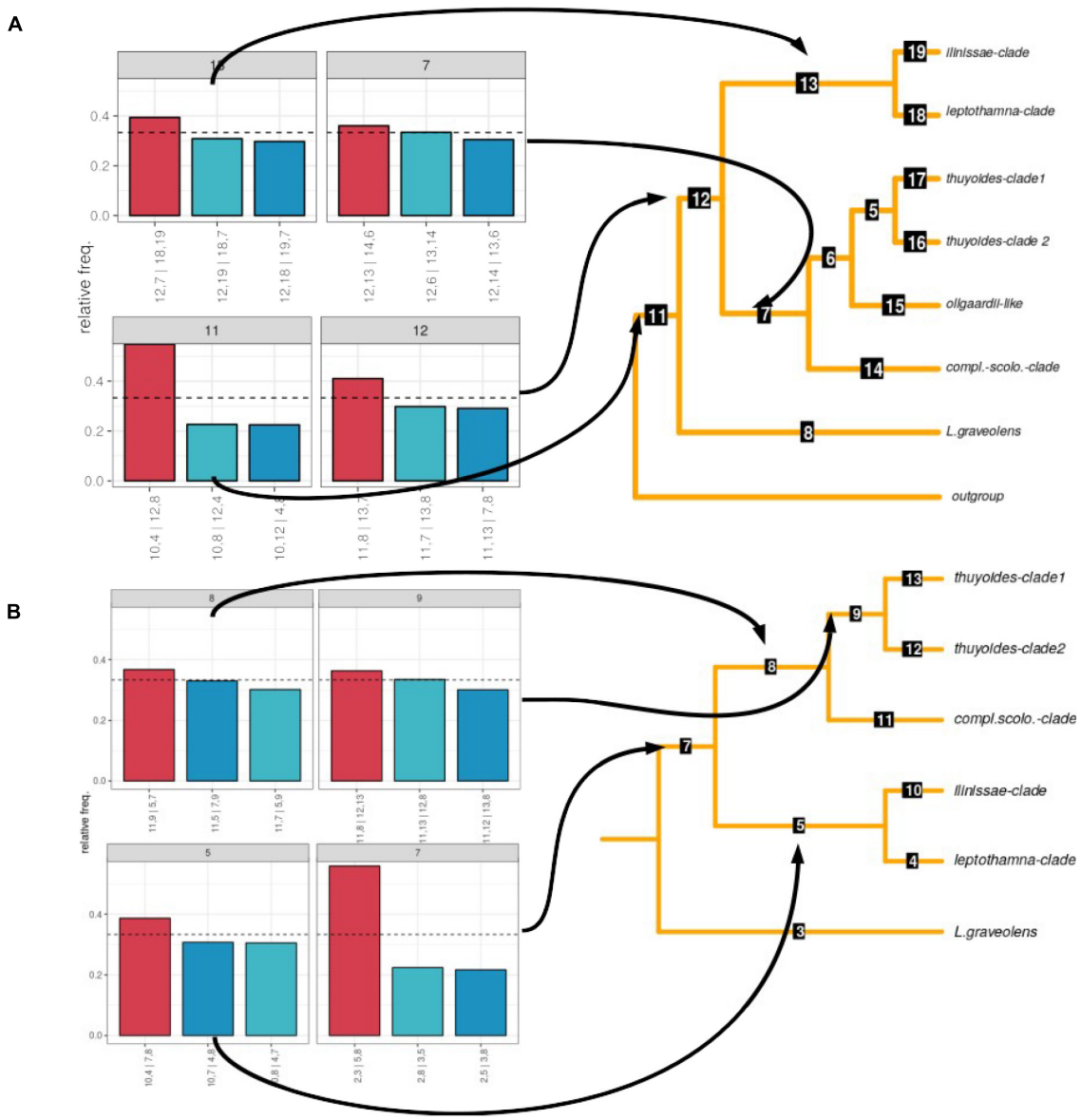


FIGURE 3 | Frequency of alternative topologies supported by gene trees before (dataset 1; **A**) and after removing samples that showed gene flow according to Dsuite (dataset 2; **B**). The relative frequencies of the topologies are shown on the left. On the right, the main topologies are shown that are reduced to clades; the numbers on the branches indicate node numbers.

While the samples identified as discordant varied between the different methods, they all belonged to the Axilliary-clade (**Table 2**). The Dsuite analyses indicated that most samples showed introgression with members of the ilinissae-clade. Removing those samples (dataset 2) reduced gene tree discordance for the mrca of the Terminal- and Axilliary-clade (**Figure 3B**) and increased support for the phylogenetic backbone (**Supplementary Figure 6**).

Only after removing those samples showing introgression with samples from different clades (dataset 2) were the PhyloNet analyses able to finish. Thus, we used PhyloNet to test if all potential hybridization events were detected using Dsuite. Allowing for a maximum of three hybridization events resulted in the best model according to the AIC in PhyloNet (**Supplementary Table 4**). The structure of the five best networks within this analysis was mostly consistent, indicating that within *Loricaria* there were between one to three reticulation events, the γ-parameter ranging from 0.9 to 0.49, resulting in a hybridigenous origin of either the thuyoides-clade1 or -clade2 and/or the complanata-scolopendra-clade with an ancestor of the Terminal-clade (**Figure 4**). In four out of the five best networks, six out of 11 introgression events showed $\gamma \geq 0.3$ within at least one event per network (**Figure 4**); the remaining events showed lower levels of gene flow.

TABLE 2 | Summary of all methods that were used to detect samples showing phylogenetic discordance.

Clade	Discordant taxon	ILS	Cytonuclear discordance	NeighborNet	Introgressed clades (PhyloNet)	Degree of introgression (Dsuite sub-analysis)
Ilinissae-clade	*L. ilinissae* 11563	x				
	L. antisanensis LR_007		x			
Leptothamna-clade	*L. leptothamna-like* LR_013_XH	x				
	L. lycopodinae LR_020		x			
Complanata-scolopendra-clade	*L. scolopendra* 11610	x	–		o	
	L. complanata 14281		o			
	L. ollgaardii-like LR_011_X	x	x	o		0.07 (2)
	L. ollgaardii-like 12066		–	o		0.045 (5)
Thuyoides-clade1	*L. vermicularis* LR_012_X	x	–			
	L. azuayensis 11122	x	–			0.06 (5)
	L. scolopendra-like 1 12002	x	–	x		0.12 (3)
	L. scolopendra-like 2 11160		x			0.14 (1)
	L. thuyoides 11161	x			o	
	L. thuyoides 12444		–			0.10 (4)
	L. thuyoides 11551	x				
	L. thuyoides 11573					0.06 (2)
	L. thuyoides 12246		o			
	L. thuyoides 11523		o			
Thuyoides-clade2					o	

A dash ("–") indicates that the sample is absent from both phylogenies used for the comparison. The letter "o" indicates the placement in a different clade, the letter "x" indicates a different position within a clade. The degree of introgression using the f–branch statistic in Dsuite is only presented when values were γ > 0.05. In such cases, the number of the sub-analysis in which gene flow was detected is indicated in parentheses.

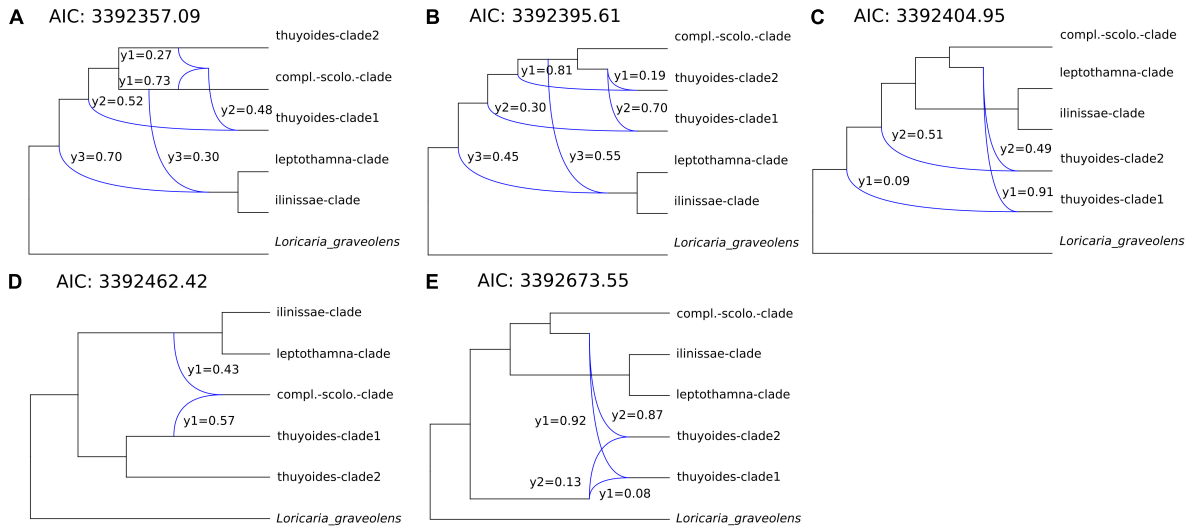


FIGURE 4 | (A–E) show the five best inferred networks with a maximum of three reticulation events according to PhyloNet using all gene trees (dataset 2). Phylogenies were reduced to clades and only nodes above a bootstrap value of 70% were considered. Gene flow between lineages is indicated by a blue line leading from two lineages to a third one. The γ-parameter indicates the degree of gene flow between different lineages. AIC values are presented and also provided in **Supplementary Table 4**.

DISCUSSION

Utility of Universal Probe Sets for Resolving Recent Rapid Radiations

Target enrichment protocols employing probe sets that may be either customized (often genus- or tribe-specific; designed to match exons across a relatively small number of species) or universal (family- or order-specific; designed to match exons across larger ranges of the tree of life) are widely used to generate hundreds of nuclear loci for samples from evolutionary lineages ranging from deep to shallow phylogenetic scales (e.g., Carlsen et al., 2018; Johnson et al., 2019; Bagley et al., 2020).

Although custom and universal probes often show a similar relative performance (Larridon et al., 2020; Shah et al., 2021; Ufimov et al., 2021), universal probes are frequently preferred for non-model organisms, for which the genomic resources, which are a prerequisite for the probe design, are often not available and too costly to generate.

In this study, we used the universal probe set Compositae1061 (Mandel et al., 2014), but customized the reference for "read fishing" before *de novo* assembly, which likely increases locus recovery and the length of recovered exons, as was shown in McLay et al. (2021) and Ufimov et al. (2021). (Universal) target enrichment kits target conserved exons that are present across a wide range of taxa. Hence, the phylogenetic signal is reduced, especially compared to introns (Folk et al., 2015; Bagley et al., 2020; Gardner et al., 2021). Nevertheless, the utility of the Compositae1061 probe set to resolve young phylogenies was shown before (Gizaw et al., 2021). By using ParalogWizard to detect paralogous loci and utilize them for phylogenetic reconstruction, we restricted our data analysis to exons, as ParalogWizard was developed for exons only. In contrast, HybPiper, the standard data analysis pipeline for Hyb-Seq data, permits to use flanking introns or supercontigs (exons plus flanking introns), potentially providing more phylogenetic informative characters (Jones et al., 2019; Ogutcen et al., 2021; Ufimov et al., 2021). However, an earlier study showed that for *Antennaria*, a close relative of *Loricaria*, very few supercontigs remained after alignment trimming and the remaining ones had low degrees of informative characters, a pattern generally prominent for Gnaphalieae (Jones et al., 2019).

While nuclear Hyb-Seq data are often able to resolve phylogenetic relationships with high support, gene tree discordance is usually high in these datasets (Jones et al., 2019; Smith et al., 2020), and the species tree topology can even be represented by only a minority of gene trees (so-called anomaly zone; Degnan and Rosenberg, 2009; Liu and Edwards, 2009; Roch and Steel, 2015). The degree of gene tree discordance tends to be lower for custom probe sets (Bagley et al., 2020; Siniscalchi et al., 2021), especially if the custom probes target longer loci compared to the universal probes (Ufimov et al., 2021). In the case of *Loricaria*, we found a high degree of gene tree discordance also in comparison to other young Asteraceae groups. However, we were able to show that using stringent filters for missing data and an elaborate analysis workflow can reduce gene tree discordance, and at it least partly explains its underlying biological processes.

Repeated rounds of WGD are common for the angiosperms (Wendel, 2015), also for the tribe Gnaphalieae and family Asteraceae (Smissen et al., 2011; Barker et al., 2016; Huang et al., 2016; Zhang et al., 2020b). We, thus, accounted for paralogy during alignment building to remove this source of gene tree discordance before addressing the effect of ILS and hybridization on discordance. Even though the Compositae1061 probe set was designed to comprise exclusively single-copy loci, a certain proportion of the loci were flagged as paralogous using HybPiper in recent works (Jones et al., 2019; Siniscalchi et al., 2019). We inferred that about 25% of the loci included paralogous copies in our dataset. The high number of paralogous loci can likely be attributed to the multiple WGD events within the family Asteraceae and the tribe Gnaphalieae in particular. As previous genome size estimates indicated that the genus lacks neopolyploids (Kolář et al., 2016), the duplicated loci are likely the result of WGDs in the tribe.

It should be noted that we do not address certain methodological artifacts as sources of gene tree discordance, namely the effect of (a) collapsing weakly supported nodes in gene trees (HPM 10), (b) removing gappy regions in the alignment (HPM 4a3), (c) selecting the most parsimony informative alignments (run manual) or (d) excluding loci showing signs of recombination (PhiPack; Bruen et al., 2006; run manual). Initial analyses showed that the effect of these artifacts on gene tree discordance was weak for our datasets (data not shown). While gene tree discordance tends to decrease with increasing data completeness (Siniscalchi et al., 2019), our stringent removal of low-quality samples and alignment filtering likely reduced the possible effect of methodological artifacts on gene tree discordance to a minimum (**Supplementary Figure 1**). Although several highly interesting samples were removed in the process of filtering for missing data, it was the only possibility to reduce gene tree estimation errors and, thus, be able to focus on the biological processes as sources of gene tree discordance (**Supplementary Figure 1**).

Phylogenetic Relationships in *Loricaria*

The genus *Loricaria* is monophyletic according to our nuclear phylogeny (**Figure 2**), while *Belloa* is nested within the earliest diverging lineage of *Loricaria* according to the plastome phylogeny (**Supplementary Figure 4**). Employing a less stringent sample filter allowed to include *Belloa* in the nuclear dataset, and in this case *Belloa* did not belong to the genus *Loricaria* in both the ASTRAL species tree and concatenated tree (data not shown). The placement of *Belloa* within *Loricaria* based on the plastome dataset provides evidence for ILS or hybridization between lineages across genera in the Gnaphalieae. This highlights that a well-sampled outgroup and good knowledge about sister lineages through broader sampling of the tribe is required to gain a better understanding of the evolution of the genus and the tribe.

The phylogeny is split into three major clades, reflecting mainly the three different sections within the genus. Strong support for species monophyly was only found for *L. graveolens*, whereas all other species were not supported to be monophyletic. Whether this is due to gene flow between species, overdescription of taxonomic species or limited phylogenetic signal for young high-altitude Andean groups for the universal Compositae1061 loci needs to be evaluated. The widespread species *L. thuyoides* is highly polyphyletic, as indicated in Kolář et al. (2016). We detected introgression between seven members of the thuyoides-clade1 and members of the ilinissae-clade using Dsuite, and another three samples showed strong cytonuclear conflict (**Supplementary Figure 4**). These results were confirmed by the PhyloNet analyses. The high degree of gene tree discordance (**Supplementary Figure 1**), while accounting for paralogy due to WGD during alignment building, suggests that ILS and hybridization played an important role in the evolution of *Loricaria*.

Signatures of Incomplete Lineage Sorting in *Loricaria*

Although the initial species tree after sample and alignment filtering (dataset 1) supported most of the major clades in the phylogeny with support > 0.95 LPP, three different gene tree topologies were almost equally likely for most nodes, indicating a substantial degree of ILS (**Figure 2A**). It needs to be noted that a comparison between the ASTRAL species tree and the concatenated tree resulted in the same set of supported clades, although within these clades some samples showed different placements (**Figure 2**). This low discordance indicates a moderate degree of ILS, as RAxML is more sensitive to ILS than ASTRAL (Mirarab et al., 2016). Collapsing nodes with low support (i.e., 10 or 30% BS) before species tree calculation did not decrease the degree of discordance (data not shown), indicating that the discordance is not due to low phylogenetic signal, but rather due to ILS.

In rapidly radiating lineages, the degree of ILS is expected to be high (Whitfield and Lockhart, 2007), due to insufficient time for alleles to coalesce. Earlier molecular dating efforts estimated a crown age of 4 Ma for *Loricaria* (Nie et al., 2016), which resulted in an approximate net diversification rate of 0.74 species per Million years [ln(N)/t; N: number of species, t: crown age; Magallon and Sanderson, 2001]. This is comparable to the high diversification rates for other plant groups from tropical high elevations in South America (Madriñán et al., 2013), as well as rates that were found in the biodiversity hotspot of the Cape Floristic Region (Pirie et al., 2016).

The degree of ILS increases with large population sizes (Slatkin and Pollack, 2006). *Loricaria* evolved during the uplift of the Andes in the northern parts of South America, where during Pleistocene glaciations the alpine belt shifted downwards (Van der Hammen, 1985; Hooghiemstra and Van der Hammen, 2004), resulting in larger potentially suitable habitats. This might have resulted in larger effective population sizes during these intervals throughout the evolution of *Loricaria*. In a future study, we will investigate the demography of *Loricaria* species using a population genomics approach.

Signatures of Hybridization in *Loricaria*

The sum of our results (cytonuclear discordance, the NeighborNet as well as the Dsuite and PhyloNet analyses) indicate that ILS alone cannot explain the high degree of discordance that we observed within *Loricaria* (**Figure 2**, **Table 2**, and **Supplementary Figure 2**). The removal of samples with introgression according to Dsuite (dataset 2) increased the support for the major split between the two main subclades in the genus from about 40% of all gene trees to above 50% (**Figure 3** and **Supplementary Figure 2**). Nevertheless, standard methods to illustrate genomic discordance did not show major improvements (based on node support and PhyParts; **Figure 2** and **Supplementary Figures 1B, 2**) between dataset 1 and dataset 2. This highlights the importance to investigate species tree hypotheses beyond support values and standard methods of measuring discordance and to thoroughly test for all potential sources of discordance, especially in highly understudied lineages.

Hybrids are unknown for the genus *Loricaria* based on morphological evidence (Cuatrecasas, 1954). In addition, previous genome size estimations of several *Loricaria* species revealed only relatively small differences (8.76–11.69 pg DNA; measurements available for *L. ilinissae, L. scolopendra, L. thuyoides*, and *L. complanata*; Kolář et al., 2016), suggesting the absence of hybridization, under the assumption that hybrids are frequently stabilized by polyploidization. Using a diverse spectrum of methods, we detected multiple hybridization events within *Loricaria*. This was not surprising given that the genus belongs to the tribe Gnaphalieae, for which many hybridization events have been reported (e.g., Smissen et al., 2011; Galbany-Casals et al., 2014; Barker et al., 2016; Huang et al., 2016; Vargas et al., 2017; Watson et al., 2020; Zhang et al., 2020b). Further, several radiating plant groups in the tropical high altitude areas of South America show hybridization (*Lachemilla*: Morales-Briones et al., 2018; *Lupinus*: Nevado et al., 2018; *Diplostephium*: Vargas et al., 2017; Espeletiinae: Cortés et al., 2018). The dynamic nature of this ecosystem with multiple range expansions and contractions during the Pleistocene (Flantua et al., 2019) may have facilitated the contact between geographically isolated species that probably did not yet exhibit strong barriers to gene flow. Using Dsuite, we identified a total of seven samples showing introgression with samples of other clades in the species tree. After removing those samples, we were able to detect one to three clades within the genus that are of hybridigenous origin according to PhyloNet (**Figure 4**). Unfortunately, PhyloNet and related methods (SNAQ; Solís-Lemus et al., 2017) are difficult to use for large datasets with hundreds of samples and a high number of hybridization events. We ran PhyloNet using the MPL algorithm, after unsuccessfully attempting to utilize the "divide and conquer" method (Zhu et al., 2019) using a maximum likelihood implementation, which did not finish (<14 days and 14 CPUs of computation) for a subset of quartets, even though the method is intended for large datasets. The subset of quartets that did not finish included to a large extent those that were subject to hybridization events based on MPL (observations during trials). While PhyloNet accounts for gene flow and ILS as a source of discordance, the type of gene flow can be the result of hybridization, introgression or horizontal gene transfer. These processes are biologically very similar and cannot be differentiated methodologically by this method. However, different degrees of gene flow between species or lineages, depicted by the γ-parameter, may hind-cast the different processes (Solís-Lemus et al., 2017), in our case supporting hybridization between early lineages within *Loricaria* (six out of 11 introgression events showed $\gamma \geq 0.3$, **Figure 4**). Nevertheless, the different degree of gene flow detected, ranging from $\gamma = 0.09$–0.49, suggests pure hybrids as well as hybridization with extensive backcrossing.

The polyphyletic nature of *L. thuyoides*, with some samples showing cytonuclear discordance, and others exhibiting introgression with members of the ilinissae-clade according to Dsuite analyses, indicates that *L. thuyoides* was subject to chloroplast capture and hybridization early in its history.

While *L. thuyoides* is described to be morphologically variable (Cuatrecasas, 1954), we could not find any morphological characters that enable samples showing introgression to be distinguished from pure samples. Chloroplast capture is the result of two species hybridizing with extensive backcrossing to one of the parents (Rieseberg and Soltis, 1991). Due to the extensive backcrossing, the nuclear signal of the hybridization event is swamped out, but the novel plastid from the hybridization event remains. The two *L. thuyoides* samples showing indication of chloroplast capture group in the plastome phylogeny with ilinissae-clade samples from close geographic proximity, a pattern common for chloroplast capture (Acosta and Premoli, 2010; Liu et al., 2020). The clades we identified to have a potentially hybridigenous origin, the scolopendra-complanata-clade and both thuyoides-clades (**Figure 4**), overlap geographically with members of the ilinissae-clade, the potential hybridization partner.

The samples of the Axilliary-clade, which showed signatures of introgression according to the Dsuite, are predominantly found in southern Ecuador, close to the Huancabamba Depression in northern Peru, which exhibits a partial interruption of the Andes by low-elevation river systems. Some works suggested that this area poses a barrier to gene flow for high altitude species (Cosacov et al., 2009; Richter et al., 2009), a pattern that cannot be confirmed by our study. Members of the two main clades, the Terminal- and Axilliary-clade, are found on both sides of this depression. Further, two of the new morphological groups were found in close proximity to this area, *L. "vermicularis" and L. "ollgaardii-like,"* and the latter was found to be of hybridigenous origin. Due to the shifts of the alpine belt during the Pleistocene glaciation and deglaciation cycles (Flantua et al., 2019), populations of the different clades likely came into contact in the Huancabamba Depression, which might have facilitated hybridization (L. *"ollgaardii-like," L. "scolopendra-like" 1 and 2*) as well as speciation (*L. "vermicularis"*). As such, for species of *Loricaria* that are exclusively found in the páramo ecosystem, the Huancabamba area does not seem to be a barrier to gene flow. The Huancabamba Depression has also been identified as a center of diversity for montane species (Mutke et al., 2014; Quintana et al., 2017). We cannot confirm, however, if this secondary contact was also facilitated in the southern part of the Huancabamba Depression, as we lack good sampling from the northern parts of Peru.

Despite evidence for hybridization within *Loricaria*, the exact parents and clades subject to the hybridization events could not be determined. Due to our simplification in the PhyloNet analysis, by mapping samples to supported clades, we could not differentiate if only some of the species in the clade or the clades as a whole were subject to the hybridization events. To elucidate which species are of hybrid origin and which parental species gave rise to these hybrids, further sampling in the area is needed as well as population-level analyses.

Hundreds of loci and thorough testing of potential causes of discordance provided a better understanding of the evolution of the genus. And yet, while nowadays detecting hybrids using genomic data is easier than during Sanger sequencing times, the lack of knowledge about lineages and missing taxonomic expertise in young radiations complicate our understanding of their evolution.

DATA AVAILABILITY STATEMENT

The data generated for this study can be found in Genbank SRA under Bioproject number: PRJNA777419 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA777419).

AUTHOR CONTRIBUTIONS

RS, MK, PS, and FK conceived and designed the research. PS performed the fieldwork and curated the plant material. MK processed the data, performed the phylogenetic analyses, and led the manuscript preparation. RS supervised the analyses and improved the manuscript. RS and FK facilitated the project by logistic and infrastructure support. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Czech Science Foundation GAČR project No. 20-10878S to RS and FK. It was also supported by long-term research development project No. RVO 67985939 of the Czech Academy of Sciences.

ACKNOWLEDGMENTS

We thank the reviewers for their very valuable comments and suggestions for improvement. We thank Juan Manuel Gorospe for his help with figures and Katy Jones for correcting language. We also thank Luciana Salomón, Tomáš Fér, and Juan Manuel Gorospe for discussions about disentangling phylogenetic discordance. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.765719/full#supplementary-material

REFERENCES

Acosta, M. C., and Premoli, A. C. (2010). Evidence of chloroplast capture in South American *Nothofagus* (subgenus *Nothofagus*, Nothofagaceae). *Mol. Phylogenet. Evol.* 54, 235–242. doi: 10.1016/j.ympev.2009.08.008

Anderson, E., and Hubricht, L. (1938). Hybridization in *Tradescantia*. III. The evidence for introgressive hybridization. *Am. J. Bot.* 25, 396–402. doi: 10.2307/2436413

Bagheri, A., Maassoumi, A. A., Rahiminejad, M. R., Brassac, J., and Blattner, F. R. (2017). Molecular phylogeny and divergence times of *Astragalus* section Hymenostegis: an analysis of a rapidly diversifying species group in Fabaceae. *Sci. Rep.* 7:14033. doi: 10.1038/s41598-017-14614-3

Bagley, J. C., Uribe-Convers, S., Carlsen, M. M., and Muchhala, N. (2020). Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: neotropical *Burmeistera* bellflowers as a case study. *Mol. Phylogenet. Evol.* 152:106769. doi: 10.1016/j.ympev.2020.106769

Barker, M. S., Li, Z., Kidder, T. I., Reardon, C. R., Lai, Z., Oliveira, L. O., et al. (2016). Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. *Am. J. Bot.* 103, 1203–1211. doi: 10.3732/ajb.1600113

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

Bruen, T. C., Philippe, H., and Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172, 2665–2681. doi: 10.1534/genetics.105.048975

Bryant, D., and Moulton, V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* 21, 255–265. doi: 10.1093/molbev/msh018

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner (No. LBNL-7065E)*. Berkeley, CA: Lawrence Berkeley National Lab (LBNL).

Carlsen, M. M., Fér, T., Schmickl, R., Leong-Škorničková, J., Newman, M., and Kress, W. J. (2018). Resolving the rapid plant radiation of early diverging lineages in the tropical Zingiberales: pushing the limits of genomic data. *Mol. Phylogenet. Evol.* 128, 55–68. doi: 10.1016/j.ympev.2018.07.020

Constantinides, B., and Robertson, D. L. (2017). Kindel: indel-aware consensus for nucleotide sequence alignments. *J. Open Source Softw.* 2:282.

Contreras-Ortiz, N., Atchison, G. W., Hughes, C. E., and Madriñán, S. (2018). Convergent evolution of high elevation plant growth forms and geographically structured variation in Andean *Lupinus* (Fabaceae). *Botanical J. Linnean Soc.* 187, 118–136. doi: 10.1093/botlinnean/box095

Cortés, A. J., Garzón, L. N., Valencia, J. B., and Madriñán, S. (2018). On the causes of rapid diversification in the páramos: isolation by ecology and genomic divergence in *Espeletia*. *Front. Plant Sci.* 9:1700. doi: 10.3389/fpls.2018.01700

Cosacov, A., Sérsic, A. N., Sosa, V., De-Nova, J. A., Nylinder, S., and Cocucci, A. A. (2009). New insights into the phylogenetic relationships, character evolution, and phytogeographic patterns of *Calceolaria* (Calceolariaceae). *Am. J. Bot.* 96, 2240–2255. doi: 10.3732/ajb.0900165

Cuatrecasas, J. (1954). Synopsis der Gattung *Loricaria* Wedd. *Feddes Repert* 56, 149–172. doi: 10.1002/fedr.19540560204

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Degnan, J. H. (2018). Modeling hybridization under the network multispecies coalescent. *Syst. Biol.* 67, 786–799. doi: 10.1093/sysbio/syy040

Degnan, J. H., and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340. doi: 10.1016/j.tree.2009.01.009

Diazgranados, M., and Barber, J. C. (2017). Geography shapes the phylogeny of frailejones (Espeletiinae Cuatrec., Asteraceae): a remarkable example of recent rapid radiation in sky islands. *PeerJ* 5:e2968. doi: 10.7717/peerj.2968

Dillon, M. O., and Sagastegui Alva, A. (1986). New species and status changes in Andean Inuleae (Asteraceae). *Phytologia* 59, 227–233. doi: 10.5962/bhl.part.2767

Drummond, C. S., Eastwood, R. J., Miotto, S. T. S., and Hughes, C. E. (2012). Multiple continental radiations and correlates of diversification in *Lupinus* (Leguminosae): testing for key innovation with incomplete taxon sampling. *Syst. Biol.* 61, 443–460.

Escudero, M., Nieto Feliner, G., Pokorny, L., Spalink, D., and Viruel, J. (2020). Editorial: phylogenomic approaches to deal with particularly challenging plant lineages. *Front. Plant Sci.* 11:591762. doi: 10.3389/fpls.2020.591762

Esselstyn, J. A., Oliveros, C. H., Swanson, M. T., and Faircloth, B. C. (2017). Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. *Genome Biol. Evol.* 9, 2308–2321. doi: 10.1093/gbe/evx168

Fér, T., and Schmickl, R. E. (2018). HybPhyloMaker: target enrichment data analysis from raw reads to species trees. *Evol. Bioinform.* 14:1176934317742613. doi: 10.1177/1176934317742613

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Biol.* 19, 99–113. doi: 10.2307/2412448

Flantua, S. G. A., O'Dea, A., Onstein, R. E., Giraldo, C., and Hooghiemstra, H. (2019). The flickering connectivity system of the north Andean páramos. *J. Biogeogr.* 46, 1808–1825. doi: 10.1111/jbi.13607

Folk, R. A., Mandel, J. R., and Freudenstein, J. V. (2015). A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: a phylogenomic example from *Heuchera* (Saxifragaceae). *Appl. Plant Sci.* 3:1500039. doi: 10.3732/apps.1500039

Folk, R. A., Soltis, P. S., Soltis, D. E., and Guralnick, R. (2018). New prospects in the detection and comparative analysis of hybridization in the tree of life. *Am. J. Bot.* 105, 364–375. doi: 10.1002/ajb2.1018

Gabaldón, T. (2008). Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.* 9:235. doi: 10.1186/gb-2008-9-10-235

Galbany-Casals, M., Andrés-Sánchez, S., Garcia-Jacas, N., Susanna, A., Rico, E., and Martínez-Ortega, M. M. (2010). How many of Cassini anagrams should there be? Molecular systematics and phylogenetic relationships in the Filago group (Asteraceae, Gnaphalieae), with special focus on the genus Filago. *Taxon* 59, 1671–1689. doi: 10.1002/tax.596003

Galbany-Casals, M., Unwin, M., Smissen, R. D., Susanna, A., and Bayer, R. J. (2014). Phylogenetic relationships in *Helichrysum* (Compositae: Gnaphalieae) and related genera: incongruence between nuclear and plastid phylogenies, biogeographic and morphological patterns, and implications for generic delimitation. *Taxon* 63, 608–624.

Gardner, E. M., Johnson, M. G., Pereira, J. T., Puad, A. S. A., Arifiani, D., Wickett, N. J., et al. (2021). Paralogs and off-target sequences improve phylogenetic resolution in a densely sampled study of the breadfruit genus (*Artocarpus*, Moraceae). *Syst. Biol.* 70, 558–575. doi: 10.1093/sysbio/syaa073

Givnish, T. J. (2015). Adaptive radiation versus 'radiation' and 'explosive diversification': why conceptual distinctions are fundamental to understanding evolution. *New Phytol.* 207, 297–303.

Gizaw, A., Gorospe, J. M., Kandziora, M., Chala, D., Gustafsson, L., Zinaw, A., et al. (2021). Afro-alpine flagships revisited II: elucidating the evolutionary relationships and species boundaries in the giant senecios *Dendrosenecio*. *Alpine Bot.* 1–17. doi: 10.1007/s00035-021-00268-5

Grunewald, S., Spillner, A., Bastkowski, S., Bogershausen, A., and Moulton, V. (2013). SuperQ: computing supernetworks from quartets. *IEEE/ACM Trans. Computat. Biol. Bioinform.* 10, 151–160. doi: 10.1109/TCBB.2013.8

Hind, D. J. N. (2004). A new species of *Loricaria* (Compositae: Inuleae sensu lato) from Ecuador. *Kew Bull.* 59:541. doi: 10.2307/4110908

Hooghiemstra, H., and Van der Hammen, T. (2004). Quaternary Ice-Age dynamics in the Colombian Andes: developing an understanding of our legacy. *Philosop. Trans. R. Soc. B* 359, 173–181. doi: 10.1098/rstb.2003.1420

Huang, C.-H., Zhang, C., Liu, M., Hu, Y., Gao, T., Qi, J., et al. (2016). Multiple polyploidization events across asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Mol. Biol. Evol.* 33, 2820–2835. doi: 10.1093/molbev/msw157

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14, 68–73. doi: 10.1093/bioinformatics/14.1.68

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Jiang, X., Edwards, S. V., and Liu, L. (2020). The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. *Syst. Biol.* 69, 795–812. doi: 10.1093/sysbio/syaa008

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., et al. (2016). HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* 4:1600016. doi: 10.3732/apps.1600016

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., et al. (2019). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-Medoids clustering. *Syst. Biol.* 68, 594–606. doi: 10.1093/sysbio/syy086

Jones, K. E., Fér, T., Schmickl, R. E., Dikow, R. B., Funk, V. A., Herrando-Moraira, S., et al. (2019). An empirical assessment of a single family-wide hybrid capture locus set at multiple evolutionary timescales in Asteraceae. *Appl. Plant Sci.* 7:e11295. doi: 10.1002/aps3.11295

Kamneva, O. K., Syring, J., Liston, A., and Rosenberg, N. A. (2017). Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC Evol. Biol.* 17:180. doi: 10.1186/s12862-017-1019-7

Kandziora, M., Kadereit, J. W., and Gehrke, B. (2016). Frequent colonization and little in situ speciation in *Senecio* in the tropical alpine-like islands of eastern Africa. *Am. J. Bot.* 103, 1483–1498. doi: 10.3732/ajb.1600210

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511–518. doi: 10.1093/nar/gki198

Kolář, F., Dušková, E., and Sklenář, P. (2016). Niche shifts and range expansions along cordilleras drove diversification in a high-elevation endemic plant genus in the tropical Andes. *Mol. Ecol.* 25, 4593–4610. doi: 10.1111/mec.13788

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455. doi: 10.1093/bioinformatics/btz305

Larridon, I., Villaverde, T., Zuntini, A. R., Pokorny, L., Brewer, G. E., Epitawalage, N., et al. (2020). Tackling rapid radiations with targeted sequencing. *Front. Plant Sci.* 10:1655. doi: 10.3389/fpls.2019.01655

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., and Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). *New Phytol.* 221, 515–526. doi: 10.1111/nph.15386

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinform.* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinform.* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Liu, B.-B., Campbell, C. S., Hong, D.-Y., and Wen, J. (2020). Phylogenetic relationships and chloroplast capture in the *Amelanchier-Malacomeles-Peraphyllum* clade (Maleae, Rosaceae): evidence from chloroplast genome and nuclear ribosomal DNA data using genome skimming. *Mol. Phylogenet. Evol.* 147:106784. doi: 10.1016/j.ympev.2020.106784

Liu, L., and Edwards, S. V. (2009). Phylogenetic analysis in the anomaly zone. *Syst. Biol.* 58, 452–460. doi: 10.1093/sysbio/syp034

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.* 46, 523–536.

Madriñán, S., Cortés, A. J., and Richardson, J. E. (2013). Páramo is the world's fastest evolving and coolest biodiversity hotspot. *Front. Genet.* 4:192. doi: 10.3389/fgene.2013.00192

Magallon, S., and Sanderson, M. J. (2001). Absolute diversification rates in angiosperm clades. *Evolution* 55, 1762–1780.

Malinsky, M., Matschiner, M., and Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Mol. Ecol. Resour.* 21, 584–595. doi: 10.1111/1755-0998.13265

Mandel, J. R., Dikow, R. B., Funk, V. A., Masalia, R. R., Staton, S. E., Kozik, A., et al. (2014). A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Appl. Plant Sci.* 2:1300085. doi: 10.3732/apps.1300085

Mandel, J. R., Dikow, R. B., Siniscalchi, C. M., Thapa, R., Watson, L. E., and Funk, V. A. (2019). A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. *Proc. Natl. Acad. Sci. U.S.A.* 116, 14083–14088. doi: 10.1073/pnas.1903871116

McLay, T. G. B., Birch, J. L., Gunn, B. F., Ning, W., Tate, J. A., Nauheimer, L., et al. (2021). New targets acquired: improving locus recovery from the Angiosperms353 probe set. *Appl. Plant Sci.* 9:10.1002/as3.11420. doi: 10.1002/aps3.11420

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., and Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat. Commun.* 8:14363. doi: 10.1038/ncomms14363

Mirarab, S., Bayzid, M. S., and Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst. Biol.* 65, 366–380.

Molloy, E. K., and Warnow, T. (2018). To include or not to include: the impact of gene filtering on species tree estimation methods. *Syst. Biol.* 67, 285–303. doi: 10.1093/sysbio/syx077

Morales-Briones, D. F., Kadereit, G., Tefarikis, D. T., Moore, M. J., Smith, S. A., Brockington, S. F., et al. (2021). Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae s.l. *Syst. Biol.* 70, 219–235. doi: 10.1093/sysbio/syaa066

Morales-Briones, D. F., Liston, A., and Tank, D. C. (2018). Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytol.* 218, 1668–1684. doi: 10.1111/nph.15099

Mutke, J., Jacobs, R., Meyers, K., Henning, T., and Weigend, M. (2014). Diversity patterns of selected Andean plant groups correspond to topography and habitat dynamics, not orogeny. *Front. Genet.* 5:351. doi: 10.3389/fgene.2014.00351

Nevado, B., Contreras-Ortiz, N., Hughes, C., and Filatov, D. A. (2018). Pleistocene glacial cycles drive isolation, gene flow and speciation in the high-elevation Andes. *New Phytologist* 219, 779–793. doi: 10.1111/nph.15243

Nie, Z.-L., Funk, V. A., Meng, Y., Deng, T., Sun, H., and Wen, J. (2016). Recent assembly of the global herbaceous flora: evidence from the paper daisies (Asteraceae: Gnaphalieae). *New Phytologist* 209, 1795–1806. doi: 10.1111/nph.13740

Nute, M., Chou, J., Molloy, E. K., and Warnow, T. (2018). The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics* 19:286. doi: 10.1186/s12864-018-4619-8

Ogutcen, E., Christe, C., Nishii, K., Salamin, N., Möller, M., and Perret, M. (2021). Phylogenomics of Gesneriaceae using targeted capture of nuclear genes. *Mol. Phylogenet. Evol.* 157:107068. doi: 10.1016/j.ympev.2021.107068

Ottenlips, M. V., Mansfield, D. H., Buerki, S., Feist, M. A. E., Downie, S. R., Dodsworth, S., et al. (2021). Resolving species boundaries in a recent radiation with the Angiosperms353 probe set: the *Lomatium packardiae/L. anomalum* clade of the *L. triternatum* (Apiaceae) complex. *Am. J. Bot.* 108, 1217–1233. doi: 10.1002/ajb2.1676

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., et al. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* 2:e000056. doi: 10.1099/mgen.0.000056

Panero, J. L., and Crozier, B. S. (2016). Macroevolutionary dynamics in the early diversification of Asteraceae. *Mol. Phylogenet. Evol.* 99, 116–132. doi: 10.1016/j.ympev.2016.03.007

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093. doi: 10.1534/genetics.112.145037

Pirie, M. D., Oliver, E. G. H., Mugrabi de Kuppler, A., Gehrke, B., Le Maitre, N. C., Kandziora, M., et al. (2016). The biodiversity hotspot as evolutionary hot-bed: spectacular radiation of *Erica* in the Cape Floristic Region. *BMC Evol. Biol.* 16:190. doi: 10.1186/s12862-016-0764-3

Quintana, C., Pennington, R. T., Ulloa, C. U., and Balslev, H. (2017). Biogeographic barriers in the Andes: is the Amotape—Huancabamba zone a dispersal barrier for dry forest plants? *Ann. Missouri Botanical Garden* 102, 542–550. doi: 10.3417/D-17-00003A

Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., et al. (2017). Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Syst. Biol.* 66, 857–879. doi: 10.1093/sysbio/syx041

Richter, M., Diertl, K.-H., Emck, P., Peters, T., and Beck, E. (2009). Reasons for an outstanding plant diversity in the tropical Andes of Southern Ecuador. *Landscape Online* 12, 1–35. doi: 10.3097/LO.200912

Rieseberg, L. H., and Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trends Plants* 5, 65–84.

Roch, S., and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Popul. Biol.* 100, 56–62. doi: 10.1016/j.tpb.2014.12.005

Sayyari, E., and Mirarab, S. (2018). Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes* 9:132.

Sayyari, E., Whitfield, J. B., and Mirarab, S. (2018). DiscoVista: interpretable visualizations of gene tree discordance. *Mol. Phylogenet. Evol.* 122, 110–115. doi: 10.1016/j.ympev.2018.01.019

Schluter, D. (2000). *The Ecology of Adaptive Radiation*. Oxford: Oxford University Press.

Shah, T., Schneider, J. V., Zizka, G., Maurin, O., Baker, W., Forest, F., et al. (2021). Joining forces in Ochnaceae phylogenomics: a tale of two targeted sequencing probe kits. *Am. J. Bot.* 108, 1201–1216. doi: 10.1002/ajb2.1682

Siniscalchi, C. M., Hidalgo, O., Palazzesi, L., Pellicer, J., Pokorny, L., Maurin, O., et al. (2021). Lineage-specific vs. universal: a comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family. *Appl. Plant Sci.* 9:10.1002/as3.11422. doi: 10.1002/aps3.11422

Siniscalchi, C. M., Loeuille, B., Funk, V. A., Mandel, J. R., and Pirani, J. R. (2019). Phylogenomics yields new insight into relationships within Vernonieae (Asteraceae). *Front. Plant Sci.* 10:1224. doi: 10.3389/fpls.2019.01224

Sklenář, P., Dušková, E., and Balslev, H. (2011). Tropical and Temperate: evolutionary history of Páramo Flora. *Bot. Rev.* 77, 71–108. doi: 10.1007/s12229-010-9061-9

Slatkin, M., and Pollack, J. L. (2006). The concordance of gene trees and species trees at two linked loci. *Genetics* 172, 1979–1984. doi: 10.1534/genetics.105.049593

Smissen, R. D., Galbany-Casals, M., and Breitwieser, I. (2011). Ancient allopolyploidy in the everlasting daisies (Asteraceae: Gnaphalieae): complex relationships among extant clades. *Taxon* 60, 649–662.

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* 15:150. doi: 10.1186/s12862-015-0423-0

Smith, S. A., Walker-Hale, N., and Walker, J. F. (2020). Intragenic conflict in phylogenomic data sets. *Mol. Biol. Evol.* 37, 3380–3388. doi: 10.1093/molbev/msaa170

Solís-Lemus, C., Bastide, P., and Ané, C. (2017). PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.* 34, 3292–3298. doi: 10.1093/molbev/msx235

Song, S., Liu, L., Edwards, S. V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14942–14947. doi: 10.1073/pnas.1211733109

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

Sun, M., Soltis, D. E., Soltis, P. S., Zhu, X., Burleigh, J. G., and Chen, Z. (2015). Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Mol. Phylogenet. Evol.* 83, 156–166. doi: 10.1016/j.ympev.2014.11.003

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

Thomas, A. E., Igea, J., Meudt, H. M., Albach, D. C., Lee, W. G., and Tanentzap, A. J. (2021). Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand *Veronica*. *Am. J. Bot.* 108, 1289–1306. doi: 10.1002/ajb2.1678

Townsend, J. P. (2007). Profiling phylogenetic informativeness. *Syst. Biol.* 56, 222–231. doi: 10.1080/10635150701311362

Ufimov, R., Zeisek, V., Píšová, S., Baker, W. J., Fér, T., Loo, M., et al. (2021). Relative performance of customized and universal probe sets in target enrichment: a case study in subtribe Malinae. *Appl. Plant Sci.* 9:e11442. doi: 10.1002/aps3.11442

Vachaspati, P., and Warnow, T. (2015). ASTRID: accurate species TRees from internode distances. *BMC Genomics* 16:S3. doi: 10.1186/1471-2164-16-S10-S3

Van der Hammen, T. (1985). The Plio-Pleistocene climatic record of the tropical Andes. *J. Geol. Soc.* 142, 483–489. doi: 10.1144/gsjgs.142.3.0483

Vargas, O. M., Ortiz, E. M., and 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. doi: 10.1111/nph.14530

Watson, L. E., Siniscalchi, C. M., and Mandel, J. (2020). Phylogenomics of the hyperdiverse daisy tribes: Anthemideae, Astereae, Calenduleae, Gnaphalieae, and Senecioneae. *J. Syst. Evol.* 58, 841–852. doi: 10.1111/jse.12698

Weitemier, K., Straub, S. C. K., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., et al. (2014). Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *Appl. Plant Sci.* 2:1400042. doi: 10.3732/apps.1400042

Wendel, J. F. (2015). The wondrous cycles of polyploidy in plants. *Am. J. Bot.* 102, 1753–1756. doi: 10.3732/ajb.1500320

Whitfield, J. B., and Lockhart, P. J. (2007). Deciphering ancient rapid radiations. *Trends Ecol. Evol.* 22, 258–265. doi: 10.1016/j.tree.2007.01.012

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., et al. (2017). Evolution of rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* 34, 262–281. doi: 10.1093/molbev/msw242

Yan, Z., Smith, M. L., Du, P., Hahn, M. W., and Nakhleh, L. (2021). Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Syst. Biol.* syab056. doi: 10.1093/sysbio/syab056

Yang, Y., and Smith, S. A. (2014). Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* 31, 3081–3092. doi: 10.1093/molbev/msu245

Zhang, C., Huang, C.-H., Liu, M., Hu, Y., Panero, J. L., Luebert, F., et al. (2021). Phylotranscriptomic insights into Asteraceae diversity, polyploidy, and morphological innovation. *J. Integrat. Plant Biol.* 63, 1273–1293. doi: 10.1111/jipb.13078

Zhang, C., Sayyari, E., and Mirarab, S. (2017). "ASTRAL-III: increased Scalability and Impacts of Contracting Low Support Branches," in *Comparative Genomics, Lecture Notes in Computer Science*, eds J. Meidanis and L. Nakhleh (Cham: Springer International Publishing), 53–75.

Zhang, C., Scornavacca, C., Molloy, E. K., and Mirarab, S. (2020a). ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* 37, 3292–3307. doi: 10.1093/molbev/msaa139

Zhang, C., Zhang, T., Luebert, F., Xiang, Y., Huang, C.-H., Hu, Y., et al. (2020b). Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Mol. Biol. Evol.* 37, 3188–3210. doi: 10.1093/molbev/msaa160

Zhu, J., Liu, X., Ogilvie, H. A., and Nakhleh, L. K. (2019). A divide-and-conquer method for scalable phylogenetic network inference from multilocus data. *Bioinformatics* 35, i370–i378. doi: 10.1093/bioinformatics/btz359

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 25 January 2022
doi: 10.3389/fpls.2021.820997

Synthesis of Nuclear and Chloroplast Data Combined With Network Analyses Supports the Polyploid Origin of the Apple Tribe and the Hybrid Origin of the Maleae—Gillenieae Clade

Richard G. J. Hodel[1], Elizabeth A. Zimmer[1], Bin-Bin Liu[1,2] and Jun Wen[1]*

[1] Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington, DC, United States, [2] State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China

OPEN ACCESS

Edited by:
Susann Wicke, Humboldt University of Berlin, Germany

Reviewed by:
Natascha D. Wagner, University of Göttingen, Germany
Ofere Francis Emeriewen, Julius Kühn Institute (JKI), Germany

***Correspondence:**
Richard G. J. Hodel
richiehodel@gmail.com

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *23 November 2021*
Accepted: *20 December 2021*
Published: *25 January 2022*

Citation:
Hodel RGJ, Zimmer EA, Liu B-B and Wen J (2022) Synthesis of Nuclear and Chloroplast Data Combined With Network Analyses Supports the Polyploid Origin of the Apple Tribe and the Hybrid Origin of the Maleae—Gillenieae Clade. Front. Plant Sci. 12:820997. doi: 10.3389/fpls.2021.820997

Plant biologists have debated the evolutionary origin of the apple tribe (Maleae; Rosaceae) for over a century. The "wide-hybridization hypothesis" posits that the pome-bearing members of Maleae (base chromosome number $x = 17$) resulted from a hybridization and/or allopolyploid event between progenitors of other tribes in the subfamily Amygdaloideae with $x = 8$ and $x = 9$, respectively. An alternative "spiraeoid hypothesis" proposed that the $x = 17$ of Maleae arose via the genome doubling of $x = 9$ ancestors to $x = 18$, and subsequent aneuploidy resulting in $x = 17$. We use publicly available genomic data—448 nuclear genes and complete plastomes—from 27 species representing all major tribes within the Amygdaloideae to investigate evolutionary relationships within the subfamily containing the apple tribe. Specifically, we use network analyses and multi-labeled trees to test the competing wide-hybridization and spiraeoid hypotheses. Hybridization occurred between an ancestor of the tribe Spiraeeae ($x = 9$) and an ancestor of the clade Sorbarieae ($x = 9$) + Exochordeae ($x = 8$) + Kerrieae ($x = 9$), giving rise to the clade Gillenieae ($x = 9$) + Maleae ($x = 17$). The ancestor of the Maleae + Gillenieae arose via hybridization between distantly related tribes in the Amygdaloideae (i.e., supporting the wide hybridization hypothesis). However, some evidence supports an aspect of the spiraeoid hypothesis—the ancestors involved in the hybridization event were likely both $x = 9$, so genome doubling was followed by aneuploidy to result in $x = 17$ observed in Maleae. By synthesizing existing genomic data with novel analyses, we resolve the nearly century-old mystery regarding the origin of the apple tribe. Our results also indicate that nuclear gene tree-species tree conflict and/or cytonuclear conflict are pervasive at several other nodes in subfamily Amygdaloideae of Rosaceae.

Keywords: allopolyploidy, ancient hybridization, cytonuclear conflict, genome doubling, phylogenetic networks, phylogenomics, reticulate evolution

INTRODUCTION

Throughout the Rosaceae, there is pervasive conflict between phylogenetic relationships inferred using the nuclear vs. chloroplast genomes. Among major lineages of the Rosaceae, variation in chromosome number is prevalent, and there have been frequent whole genome duplications in the family. Many lineages of the Rosaceae contain economically important species; the Maleae, with over 1,000 species, includes commercially important fruit crops, such as apples and pears, as well as many ornamentals. In addition to apples and pears, the subfamily Amygdaloideae contains many other important species such as cherries, almonds, peaches, apricots, and plums. The branching order among the three subfamilies of the Rosaceae—Amygdaloideae, Dryadoideae, and Rosoideae—is uncertain. Nuclear data indicate that the Dryadoideae are sister to the Amygdaloideae + Rosoideae (Xiang et al., 2017), whereas phylogenetic relationships reconstructed using plastome data have still not conclusively resolved the branching order. Recent analyses inferred that the Rosoideae are sister to Amygdaloideae + Dryadoideae when using whole plastome data, or that the Amygdaloideae are sister to the Dryadoideae + Rosoideae when using whole plastomes with most ambiguous sites removed (Zhang et al., 2017). In the Amygdaloideae, the relationships between many tribes conflict when the nuclear and chloroplast topologies are compared (**Figure 1**; Xiang et al., 2017; Zhang et al., 2017). Furthermore, within the Rosaceae, many relationships between tribes were inconsistent between the nuclear and chloroplast genomes, such as the placement of all tribes within the Rosoideae except for Ulmarieae (Xiang et al., 2017; Zhang et al., 2017). Cytonuclear conflict also exists within the Rosaceae at shallower systematic scales (e.g., within the tribe Maleae; Liu et al., 2019, 2020a,b, 2021).

For nearly a century, plant biologists have debated the evolutionary origin of the apple tribe Maleae (Rosaceae; formerly Maloideae). Species in the tribe Maleae are characterized by a base chromosome number of $x = 17$ (except for $x = 15$ in *Vauquelinia* Corrêa ex Bonpl.)—distinct from other tribes in the Rosaceae, which typically are $x = 7$, 8, or 9 (Evans and Campbell, 2002). Within the Amygdaloideae, the subfamily containing the Maleae, all tribes except the Maleae are $x = 8$ or 9 (Robertson et al., 1991). Because the base chromosome number of Maleae was approximately double that of all its close relatives, early researchers investigated hypotheses of a polyploid origin of the pome-bearing members of the apple subtribe Malinae, which includes all Maleae except for three early diverging dry fruit lineages including genera *Kageneckia* Ruiz and Pav. ($x = 17$), *Lindleya* Kunth ($x = 17$) and *Vauquelinia* ($x = 15$) (Nebel, 1929; Campbell et al., 1995). Darlington and Moffett (1930) proposed hypotheses of autopolyploidy, which were quickly refuted by Sax (1931, 1932, 1933) after observing predominantly univalents in triploids during meiosis, as opposed to multivalents. Sax proposed an explanation of allopolyploidy occurring between $x = 8$ and $x = 9$ progenitors from the subfamily Spiraeoideae (now as part of Amygdaloideae; Potter et al., 2007). The "wide-hybridization hypothesis" formulated in the 1930s posits that the Malinae (base chromosome number $x = 17$) resulted from an ancient hybridization event between progenitors from other tribes in the subfamily Amygdaloideae that have $x = 8$ and $x = 9$, respectively. The "wide-hybridization" hypothesis was favored by Stebbins (1950) and was further supported by studies using isozymes decades later (Chevreau et al., 1985; Weeden and Lamb, 1987).

An alternative "spiraeoid hypothesis" proposed that the 17 (or in rare cases 15) chromosomes found in Maleae arose via the genome doubling of an $x = 9$ spiraeoid ancestor to $x = 18$, and subsequent aneuploidy resulting in $x = 17$ (Goldblatt, 1976; Evans and Campbell, 2002). This hypothesis is referred to as the "spiraeoid" hypothesis because the participants in allopolyploidy were considered a member of spiraeoid taxa (Goldblatt, 1976), in particular, the ancestor of the tribe Gillenieae (Evans and Campbell, 2002), which was traditionally placed in the formerly recognized subfamily Spiraeoideae (also see Gladkova, 1972). A genetic investigation of the origin of the apple tribe using one nuclear gene (Evans and Campbell, 2002) favored the spiraeoid hypothesis while rejecting the wide-hybridization hypothesis. Their study inferred that an ancestor of the tribe Gillenieae ($x = 9$), which is sister to the Maleae, experienced genome doubling and subsequent aneuploidy. Other molecular analyses of the Rosaceae did not explicitly test hypotheses explaining the origin of the apple tribe (e.g., Potter et al., 2002, 2007). To date, the two competing hypotheses have not been tested using genomic data. Recent phylogenomic studies identified pervasive cytonuclear conflict throughout the Amygdaloideae, which contains the Maleae, suggesting that ancient hybridization and/or allopolyploidization may have impacted the diversification of this group.

The time is ripe to re-evaluate these hypotheses using analyses that consider phylogenomic data from both nuclear and chloroplast genomes, and methodologies that explicitly incorporate discordance and/or reticulation into phylogenies. As researchers obtain more DNA sequence data from both the nuclear and chloroplast genomes, it is becoming increasingly clear that cytonuclear conflict is prevalent in many plant lineages (Huang et al., 2014; Bruun-Lund et al., 2017; Lee-Yaw et al., 2018; Hodel et al., 2021; Liu et al., 2021; Wang et al., 2021; Xu et al., 2021). Here, we limit our focus to studying and resolving cytonuclear conflict within the Amygdaloideae. Our objectives in this paper are to: (1) Test the competing wide hybridization and spiraeoid hypotheses, and investigate the role of genome doubling in the origin of the apple tribe using genomic data from the nuclear and chloroplast genomes, and (2) Characterize cytonuclear conflict within the Amygdaloideae, a clade with pervasive reticulate evolution, and identify explanations for the observed conflict. Specifically, we integrate data from Xiang et al. (2017)—hundreds of nuclear genes—and plastomes from Zhang et al. (2017), supplemented by chloroplast sequence data from NCBI, to investigate pervasive cytonuclear conflict within the Amygdaloideae that may provide insights into the evolutionary origin of the apple tribe.

MATERIALS AND METHODS

Dataset Construction

Subfamily Amygdaloideae contains approximately 1,500 species organized into nine tribes (**Figure 1**). Some tribes, such as Maleae and Amygdaleae are represented by hundreds of species, whereas others such as Gillenieae and Lyonothamneae each contain a single genus. We selected representatives from each tribe, as well as species from the other Rosaceae subfamilies Dryadoideae and Rosoideae, with the goal of obtaining a representative sampling of Amygdaloideae tribes while limiting the number of taxa included so that certain analyses (i.e., phylogenetic networks) would be computationally feasible. First, we downloaded the 148-taxa alignments of 882 nuclear genes from Xiang et al. (2017) from TreeBASE (study ID = 19726). Briefly, Xiang et al. (2017) isolated RNA from young leaf, floral bud, or fruit tissue, performed transcriptome sequencing, and identified putative low copy candidate orthologous genes to use in phylogenentic analyses. The publicly available alignments consisted of consensus sequences from the candidate orthologous genes. As the authors of Xiang et al. (2017) note, a large proportion of these 882 nuclear genes are suspected hidden paralogs, and they used several paralog filtering steps. Xiang et al. (2017) primarily used smaller filtered subsets of genes (571, 444, 256, and 113 genes) in phylogenomic analyses. In the present study, phylogenies were first constructed using all 148 taxa to identify putative paralogous gene trees. We inferred each of 882 gene trees from the sequence alignments using RAxML v8.2.11 (Stamatakis, 2014) with the GTRGAMMA model of evolution, 20 independent ML searches, and 100 bootstrap replicates. For consistency with Xiang et al. (2017), we screened all gene trees using TreSpEx (Struck, 2014) with the *a priori* paralogy screening function with a bootstrap threshold of 95—with two masking filters—first using established ordinal relationships, and then using subfamilial relationships. The ordinal and subfamilial filters were used in Xiang et al. (2017) to remove suspected hidden paralogs, so we used this strategy for consistency. Our TreSpEx paralog trimming left 448 putative orthologs out of 882. When we investigated including paralogs in our analyses (i.e., using all 882 genes), our species tree topology did not match the dominant topology presented in Xiang et al. (2017). Therefore, we proceeded using our 448 gene set, which did match the dominant topology from Xiang et al. (2017). We trimmed taxa from the 448-gene alignments using the "pxrmt" command in *phyx* (phylogenetic tools for unix; Brown et al., 2017) to reduce the data matrix down to 27 species. We included at least one species from each of the nine tribes in the Amygdaloideae and two species each from the Rosoideae and Dryadoideae, as well as one outgroup species, *Ziziphus jujuba* Mill. (Rhamnaceae). The trimming of taxa was done to facilitate downstream analyses (i.e., network analyses implemented in SNaQ) that become computationally intractable when larger numbers of taxa (i.e., > 30) are included. Whenever possible, we selected species represented by both nuclear data (from Xiang et al., 2017) and complete plastomes (from Zhang et al., 2017).

For the species not represented by plastome data in Zhang et al. (2017), we downloaded complete plastomes from NCBI for all species except *Physocarpus opulifolius* (L.) Maxim. (**Table 1**), which was represented in the nuclear data from Xiang et al. (2017) but not in the plastome data from Zhang et al. (2017). For *P. opulifolius*, we downloaded RNA-Seq reads from NCBI (accession number ERR2040427; **Table 1**) and used FastPlast[1] to *de novo* assemble reads into contigs. The contigs were mapped to a reference plastome [*Malus domestica* (Suckow) Borkh., accession number: MK434916.1; **Table 1**] to complete the assembly. Because this species was the only taxon without a complete plastome sequence, we included two additional *Physocarpus* (Cambess.) Raf. plastomes from Zhang et al. (2017), labeled by the authors as *Physocarpus sp. A* and *Physocarpus sp. B*, in our preliminary chloroplast phylogenetic analyses to verify that the phylogenetic position of our newly assembled plastome was as expected. MAFFT (Katoh and Standley, 2013) was used to align the plastomes with settings "--maxiterate 5000 --localpair --adjustdirectionaccurately." Resulting alignments were trimmed using TrimAl with the "-automated1" heuristic. The "pxclsq" command in *phyx* was separately used to filter the alignment based on either 20, 30, 40, 50, or 60% column occupancy required. We compared the phylogenetic trees resulting from all alignments, and after determining there was no change in topology, we used the TrimAl-trimmed tree in subsequent plastome phylogenetic analyses.

We assessed the phylogenetic relationships among the 27 species using RAxML and ASTRAL to ensure that the nuclear topology reflected the relationships from Xiang et al. (2017). The ML analysis was conducted in RAxML using a concatenated supermatrix of the 448 orthologues, with the GTRGAMMA model of evolution, 20 independent ML searches, and 100 bootstrap replicates. Both unpartitioned and partitioned (-q) analyses were used. The coalescent analyses were conducted in ASTRAL (Mirarab et al., 2014), a tree estimation program consistent with the coalescent, and using quartet support values to measure confidence in species relationships. The quartet support scores indicate the percentage of quartets in gene trees that are concordant with a given branch and therefore can show the amount of gene tree conflict associated with a branch. Quartet scores provide more information about uncertainty at key nodes than bootstrap scores, which can be inappropriately inflated in some phylogenomic datasets (Roycroft et al., 2020). We also used RAxML to ensure that the chloroplast relationships from Zhang et al. (2017) were recapitulated, using the GTRGAMMA model of evolution, 20 independent ML searches, and 100 bootstrap replicates. Phylogenetic trees were visualized and manipulated using IcyTree (Vaughan, 2017) and Interactive Tree of Life (Letunic and Bork, 2021). The "cophylo" function in the R package phytools (Revell, 2012) was used to visualize concordance between the nuclear and plastome phylogenies. Unless otherwise noted, all software analyses were run on the

[1]https://github.com/mrmckain/Fast-Plast.git

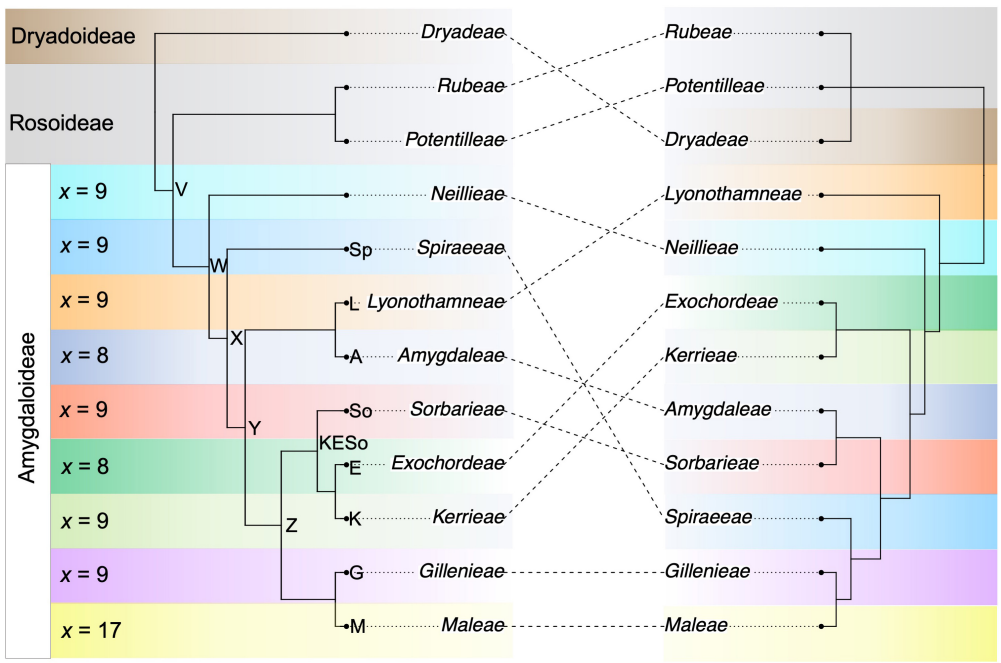


FIGURE 1 | Nuclear- and chloroplast-inferred phylogenetic relationships among tribes within the Amygdaloideae, rooted using the other two subfamilies within the Rosaceae, the Dryadoideae and Rosoideae. Conflict between the nuclear and chloroplast topologies is shown using dotted lines. The tribal and subfamilial relationships are based on topologies from Xiang et al.(2017; nuclear) and Zhang et al.(2017; chloroplast). For each tribe in the Amygdaloideae, the base chromosome number is indicated to the left of the nuclear phylogeny. Note that one genus in the Maleae is not $x = 17$ but rather is $x = 15$ (*Vauquelinia*). Select nodes in the nuclear phylogeny are labeled with letters or abbreviations to facilitate reference to these nodes in the text.

Smithsonian Institution High Performance Cluster (SI/HPC, "Hydra").[2]

Network Analyses

To assess if a reticulate tree (i.e., a phylogenetic network) better represented the nuclear gene tree data than a purely bifurcating tree, we used the program SNaQ, which is implemented in PhyloNetworks (Solís-Lemus et al., 2017). The phylogenomic network method SNaQ, which uses a pseudolikelihood method, explicitly accommodates hybridization by representing certain nodes as having received genetic material from two parental lineages with inheritance probabilities γ and 1-γ. The RAxML-inferred gene trees for all 448 orthologues were used as input and summarized using quartet concordance factors (i.e., the proportion of gene trees with a given quartet; Larget et al., 2010). In SNaQ, networks are optimized based on the branch lengths and inheritance probabilities in phylogenetic network space as measured by a pseudodeviance score. The pseudodeviance score represents a multiple of the network's log-likelihood score up to a constant where the network perfectly fits the data. Lower pseudo-deviance scores always indicate a better fit, but as h_{max} increases, the pseudodeviance score always improves (Solís-Lemus and Ané, 2016). Accordingly, the rate of change in the pseudodeviance score between h_{max} values can be used to assess the optimal h_{max} (Baudry et al., 2011). We constructed networks using h_{max} values ranging from 0 to 5. For the initial optimization ($h_{max} = 0$), the ASTRAL tree was used as a starting network with no hybridization edges, and for subsequent h_{max} values, the optimal network estimated by the preceding lower h_{max} value was used as the starting topology. We ran 10 independent searches for each h_{max} value and the optimal number of hybridization edges was assessed by plotting h_{max} against the log-likelihood score (i.e., network score) of the optimal network for each h_{max} value.

Conflict Analyses

The program *phyparts* (Smith et al., 2015) was used to assess gene tree conflict in the nuclear dataset. This program compares rooted gene trees with the rooted species tree to identify topologically concordant, discordant, and uninformative gene trees for each species tree node. Because rooted gene trees were necessary, fewer gene trees (440 out of 448) were available for this analysis due to the absence of the outgroup in some gene trees. We used a gene tree bootstrap support cutoff of 50% (-s 50); below this threshold gene trees were considered to be uninformative for a given node. A *phyparts* analysis using no bootstrap support cutoff was also run for comparison. The results of each *phyparts* analysis were visualized as piecharts on the phylogeny using

[2]https://doi.org/10.25572/SIHPC

the *phypartspiecharts.py* jupyter notebook (by Matt Johnson).[3] Nodes of interest, as identified by network analysis and the above conflict analysis, were further investigated using the "alternative relationship test" implemented in *phyckle* (Smith et al., 2020). The alternative relationship test takes as input two or more user specified bipartitions, which are used as a constraint when running RAxML to infer every gene tree from the sequence matrices. Log-likelihood scores are calculated for each gene tree and then compared to determine which topology (i.e., between the user-inputted bipartitions) is optimal for every gene tree. The number of gene trees and/or the summed difference of log-likelihood scores between the gene trees can then used to determine support for one bipartition vs. others.

Allopolyploidy Analyses

The software package GRAMPA (Thomas et al., 2017) was used to identify the parental lineages involved in a hybridization event leading to an allopolyploid lineage. GRAMPA makes use of multiply-labeled (MUL) trees, which are topologies in which selected species can appear twice, a common way of representing polyploid relationships when constrained by a bifurcating phylogeny. The algorithm implemented in GRAMPA uses least common ancestor reconciliation of gene trees and species trees (Goodman et al., 1979; Page, 1994) to place polyploidy events on a phylogeny. Branches of the species tree with disproportionately high numbers of gene duplications can be used to identify polyploidy events. The use of MUL-trees enables accurate inferences of allopolyploidy vs. autopolyploidy, because all subgenomes involved in allopolyploidy can be represented as descendants of different parental linages. Under scenarios of allopolyploidy, we would expect the homoeologs that result from an allopolyploidy event to be sister to different diploid taxa (Thomas et al., 2017). Using hypotheses from the literature, and guided by the SNaQ results, we tested the following hypotheses of allopolypoidy. We considered either the Maleae (i.e., node M; **Figure 1**) or the Gillenieae + Maleae (node G) as possible clades that were a result of allopolyploidization ("-h1" inputs). We investigated the following nodes as potential secondary parental branches ("-h2" inputs): nodes labeled A, L, Sp, S, K, E, KESo, W, X, Y, Z (**Figure 1**). If the wide hybridization hypothesis is supported, we would expect a node further removed from the Gillenieae + Maleae clade to be selected as the secondary parental branch (e.g., Sp). Conversely, if the spiraeoid-origin hypothesis is supported, we would expect the "-h2" node to be adjacent to a branch representing an ancestor of Gillenieae (e.g., Z).

[3]https://github.com/mossmatters/MJPythonNotebooks/blob/master/phypartspiecharts.py

TABLE 1 | For all 27 focal species used in our study, the NCBI accession number of the plastome sequence is listed.

Species	Chloroplast accession number	Tribe
Prunus hypoleuca	KT766059.1	Amygdaleae
Prunus mume	NC_023798.1	Amygdaleae
Prunus yedoensis	NC_026980.1	Amygdaleae
Cercocarpus montanus	KY420024.1	Dryadeae
Dryas octopetala	KY420029.1	Dryadeae
Oemleria cerasiformis	KY419923.1	Exochordeae
Prinsepia utilis	NC_021455.1	Exochordeae
Gillenia stipulata	NC_045321.1	Gillenieae
Kerria japonica	MN418902.1	Kerrieae
Rhodotypos scandens	KY419951.1	Kerrieae
Lyonothamnus floribundus	KY420005.1	Lyonothamneae
Amelanchier alnifolira	NC_045314.1	Maleae
Cydonia oblonga	MN061993.1	Maleae
Kageneckia oblonga	NC_045324.1	Maleae
Malus domestica	MK434916.1	Maleae
Rhaphiolepis indica	NC_045330.1	Maleae
Sorbus torminalis	NC_033975.1	Maleae
Vauquelinia californica	MN068269.1	Maleae
Physocarpus opulifolius	ERR2040427	Neillieae
Potentilla_freyniana	MK209638.1	Potentilleae
Rubus coreanus	NC_042715.1	Rubeae
Adenostoma fasciculatum	KY387915.1	Sorbarieae
Sorbaria sorbifolia	MN026875.1	Sorbarieae
Aruncus dioicus	MW115132.1	Spiraeeae
Holodiscus discolor	KY420032.1	Spiraeeae
Petrophytum caespitosum	KY419970.1	Spiraeeae
Ziziphus jujuba	KU351660.1	outgroup

For one species, Physocarpus opulifolius, a complete plastome sequence was not available, so we generated one from raw RNA-Seq data (accession number listed in this table); assembly details provided in text. Tribe membership for each species is indicated in the rightmost column.

Hybridization Analyses

To reconcile any differences between the phylogenetic network and MUL-tree results, we used one additional approach to test for histories of hybridization in the Amygdaloideae. The program Hybrid Detector (HyDe) uses phylogenetic invariants under a coalescent model with hybridization to infer probability of hybridization of three ingroup taxa relative to an outgroup taxon (Blischak et al., 2018). In this framework, the parameter γ represents the probability that gene trees with a hybrid population sister to parent X would arise under the parental population trees, whereas 1-γ would be the probability of a hybrid population being sister to parent Y. Based on the SNaQ results and GRAMPA results, we tested several sets of taxa for histories of hybridization in HyDe. Using the SNaQ results as a guide, we tested the hybrid status of three ingroups (Maleae + Gillenieae, Spiraeeae, Sorbarieae) relative to an outgroup (Neillieae), and based on the GRAMPA results, we tested for hybridization using three ingroups (Maleae + Gillenieae, Spiraeeae, Kerrieae + Exochordeae + Sorbarieae) and the same outgroup (Neillieae). This outgroup was chosen because it was sister to all other Amygdaloideae tribes when using nuclear data (**Figure 1**).

RESULTS

Phylogenetic Relationships

Our phylogenetic analyses recovered all subfamilies and tribes as monophyletic (**Figure 2**). In the nuclear phylogeny, the

Dryadoideae (represented by *Cercocarpus montanus* Raf. and *Dryas octopetala* L.) and Rosoideae (*Potentilla freyniana* Bornm. and *Rubus coreanus* Miq.) were successively sister to the Amygdaloideae (**Figure 2**). Within the Amygdaloideae, the Neillieae (*Physocarpus opulifolius*) and Spiraeeae [*Aruncus dioicus* (Walter) Fernald, *Holodiscus discolor* (Pursh) Maxim., and *Petrophytum caespitosum* (Nutt.) Rydb.] were successively sister to a clade containing the remaining seven tribes (**Figure 2**). The Amygdaleae [*Prunus hypoleuca* (Koehne) J.Wen, *Prunus mume* Siebold & Zucc., and *Prunus* × *yedoensis* Matsum.] and Lyonothamneae (*Lyonothamnus floribundus* Gray) then form a clade sister to the remaining five tribes. The Exochordeae (*Prinsepia utilis* Royle. and *Oemleria cerasiformis* (Torr. & Gray ex Hook. & Arn.) J.W.Landon), Kerrieae [*Kerria japonica* (L.) DC. and *Rhodotypos scandens* (Thunb.) Makino], and Sorbarieae [*Adenostoma fasciculatum* Hook. & Arn. and *Sorbaria sorbifolia* (L.) A.Braun] formed a clade that is sister to the clade comprised of Gillenieae [*Gillenia stipulata* (Muhl. ex Willd.) Nutt.] and Maleae [*Cydonia oblonga* Mill., *Sorbus torminalis* (L.) Crantz, *Malus domestica*, *Rhaphiolepis indica* (L.) Lindl. ex Ker Gawl., *Amelanchier alnifolia* (Nutt.) Nutt., *Vauquelinia californica* (Torr.) Sarg., and *Kageneckia oblonga* Ruiz & Pav.]. In the chloroplast phylogeny, the Dryadoideae and Rosoideae were sister to the Amygdaloideae (**Figure 2**). The Lyonothamneae and Neillieae were successively sister to the seven remaining tribes. Then, the Exochordeae and Kerrieae formed a clade sister to the remaining five tribes. The Amygdaleae and Sorbarieae made up a clade sister to the Spiraeeae, Gillenieae, and Maleae. Within this final clade, the Spiraeeae were sister to Maleae + Gillenieae (**Figure 2**).

The phylogeny constructed using nuclear data recapitulated results from Xiang et al. (2017) with our reduced-taxa dataset (**Figure 2** and **Supplementary Figure 1**). In the nuclear datasets, there were several topological differences between the nuclear coalescent and concatenation trees (**Figure 2** and **Supplementary Figure 1**)—differences that also existed among different datasets in Xiang et al. (2017). When comparing our nuclear phylogenies, the key difference was the placement of the Amygdaleae + Lyonothamneae clade, which was sister to the Kerrieae + Exochordeae + Sorbarieae in the concatenation trees, but in the coalescent tree was sister to these three tribes as well as the Maleae + Gillenieae (**Figure 2** and **Supplementary Figure 1**). Within the Maleae, there were also discrepancies between the coalescent tree and concatenation trees, and between the unpartitioned and partitioned concatenation trees (**Figure 2** and **Supplementary Figure 1**). In the coalescent topology, *Rhaphiolepis indica* and *Malus domestica* were respectively successively sister to *Cydonia oblonga* and *Sorbus torminalis* (**Figure 2**). However, in the unpartitioned ML phylogeny, *Malus* was sister to *Cydonia* whereas *Sorbus* was sister to *Rhaphiolepis* Lindl (**Supplementary Figure 1**). Meanwhile, in the partitioned ML tree, *Malus domestica* was sister to *Rhaphiolepis indica* and *Cydonia oblonga* was sister to *Sorbus torminalis* (**Supplementary Figure 1**). Hereafter, we use our ASTRAL topology as the nuclear topology for clarity because it matches the predominant topology presented in Xiang et al. (2017).

Our reduced-taxa plastome phylogeny matched the ML whole plastome tree topology, as opposed to the ambiguous-sites-removed tree, from Zhang et al. (2017) (**Figure 2**). For simplicity, we use this plastome tree in subsequent comparisons with the nuclear phylogeny, because the primary topological difference between plastome trees from Zhang et al. (2017) involved the branching order of subfamilies, not the relationships among Amygdaloideae tribes, which is our focus. As expected, there were numerous differences between our plastome and nuclear phylogenies (**Figure 2**) throughout the tree, including major relationships between subfamilies. In the plastome phylogeny, the Dryadoideae were sister to the Rosoideae, whereas in all nuclear trees, the Dryadoideae were sister to the Rosoideae + Amygdaloideae. There were also many differences in intertribal relationships, including virtually every tribe except the Gillenieae + Maleae (**Figure 2**). The different alignment strategies that we used for the plastome sequence alignment did not influence the inferred topology of the chloroplast phylogeny, but there was variation in the bootstrap percentages at certain nodes between the different alignments (**Supplementary Figure 2**).

Network Analyses

The SNaQ network analysis inferred that one hybridization event was optimal (**Figure 3** and **Supplementary Figure 3**). The hybridization edge indicated that the clade Gillenieae + Maleae was 57.2% sister to the Sorbarieae [represented by *Adenostoma* Hook. & Arn. and *Sorbaria* (Ser.) A.Braun], and 42.8% sister to the Spiraeeae [represented by *Aruncus* L., *Holodiscus* (K.Koch) Maxim., and *Petrophytum* (Nutt. ex Torr. & A.Gray) Rydb.; **Figure 3**]. The position of the Sorbarieae (57.2% sister to Gillenieae + Maleae) contrasted with both the nuclear topology (Sorbarieae sister to Exochordeae + Kerrieae) and the plastome topology (Sorbarieae sister to Amygdaleae) (**Figures 2**, **3**). Notably, the position of the Spiraeeae as 42.8% sister to the Gillenieae + Maleae was congruent with the plastome topology, where Spiraeeae was sister to Gillenieae + Maleae. Essentially, the major hybridization edge was similar to the nuclear topology, while the minor hybridization edge was consistent with the plastome topology (**Figures 2**, **3**). The other networks with h_{max} = 2–5 all included a hybridization edge similar to the h_{max} = 1 network (**Supplementary Figure 4**). As h_{max} increased, the network score always improved, although the very small changes in network score as h_{max} increases from 1 to 5 indicated that h_{max} = 1 was indeed the optimal network. Nevertheless, the hybrid edges in other networks can still provide valuable insights. The SNaQ network with h_{max} = 2 showed that the second hybridization edge was between *Prunus hypoleuca* of the *Maddenia* group (formerly in the genus *Maddenia* Hook. f & Thomson; Wen and Shi, 2012) and the lineage ancestral to Lyonothamneae + Amygdaleae (**Supplementary Figure 4**). This hybridization edge indicated that *Prunus hypoleuca* is 87.9% sister to the other *Prunus* L. species, and 12.1% sister to the ancestor of Lyonothamneae + Amygdaleae.

Conflict Analyses

The *phyparts* analysis indicated a wide range of gene tree conflict relative to the species tree, from virtually no conflict

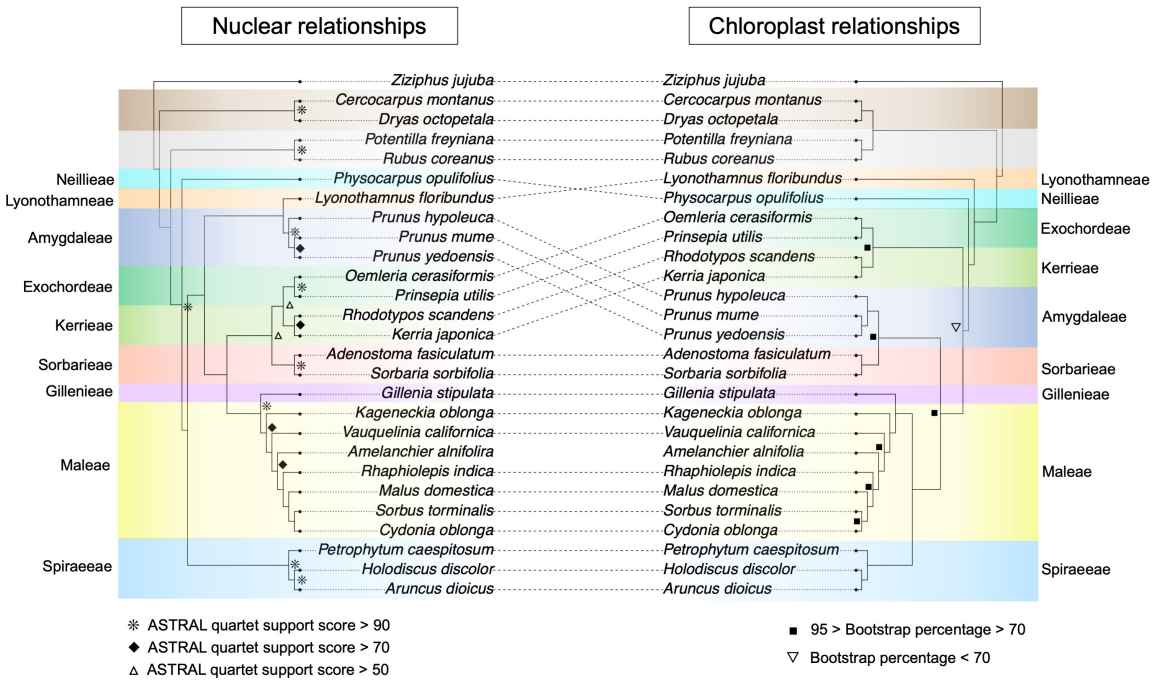


FIGURE 2 | Nuclear- and chloroplast-inferred phylogenetic relationships for all 27 focal species. Tribes/subfamilies are color-coded using the same scheme as **Figure 1**, and tribes in the Amygdaloideae are labeled. Disagreement between the chloroplast and nuclear phylogenies is indicated by dotted lines between the phylogenies. Note that in the bottom half of the phylogenies, even though the dotted lines do not intersect, there are key differences in the branching order of clades between the two phylogenies. In the nuclear phylogeny, node support was assessed using ASTRAL quartet support scores and nodes are labeled if the quartet support was greater than 50 (hollow triangles), greater than 70 (solid diamonds), or greater than 90 (asterisks). Any unlabeled nodes in the nuclear phylogeny had quartet support scores < 50. In the chloroplast phylogeny, support was measured via bootstrapping in RAxML and nodes with bootstrap scores less than 70% are labeled (inverted hollow triangles) as are nodes with less than 95% but greater than 70% (solid squares). All unlabeled nodes in the chloroplast phylogeny have 100% bootstrap support.

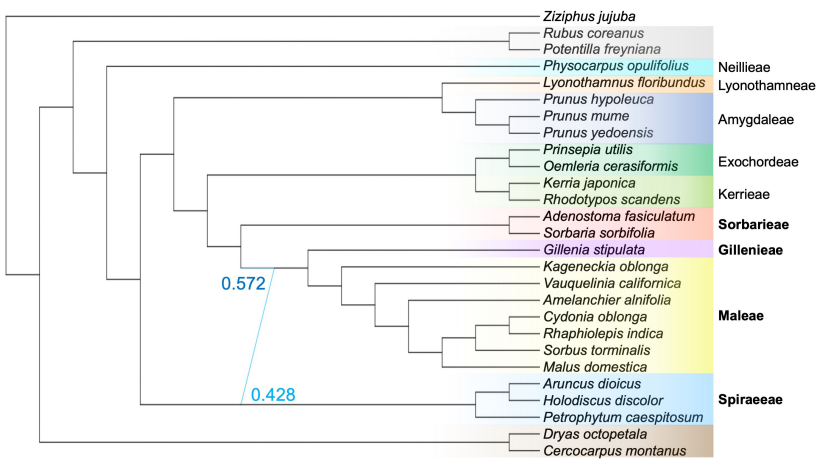


FIGURE 3 | The optimal SNaQ network using 27 species with h_{max} = 1. The SNaQ analysis inferred that the lineage *Gillenia* + Maleae is 57.2% sister to an ancestor of Sorbarieae (represented here by *Adenostoma fasiculatum* and *Sorbaria sorbifolia*), and 42.8% sister to an ancestor of Spiraeeae (represented here by *Aruncus dioicus*, *Holodiscus discolor*, *Petrophytum caespitosum*). The color coding shows tribe/subfamily membership and is identical to the scheme in **Figures 1**, **2**. Tribes in the Amygdaloideae are labeled to the right, and bold lettering indicates tribes involved in the hybridization edge.

(e.g., the node defining the Spiraeeae; node Sp1 in **Figure 4**), to pervasive conflict where nearly 10 times more genes were discordant with the species tree topology than were concordant (e.g., node Z—the node defining Gillenieae + Maleae as sister to Exochordeae + Kerrieae + Sorbarieae; **Figure 4**). The nodes with a greater proportion of gene trees in conflict with the species

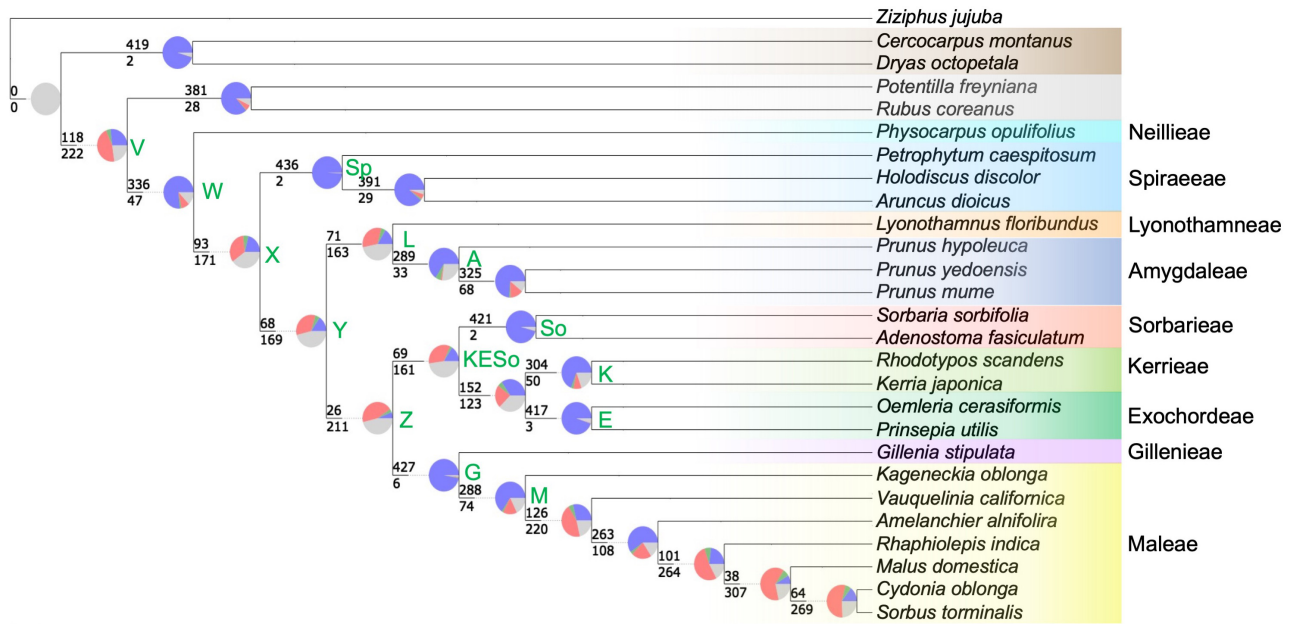


FIGURE 4 | The *phyparts* tree indicating conflict at each node of the nuclear phylogeny using the 448-gene alignment and requiring gene trees to have 50% bootstrap support to be considered informative at a given node. Because *phyparts* uses rooted gene trees, only 440 genes were included in this analysis. At each node, the pie charts indicate the proportion of homologs supporting the clade defined by the node is shown in blue, the proportion supporting the primary alternative for that clade are green, the proportion supporting all other alternatives for the clade are red, and the proportion of homologs with less than 50% bootstrap support are shown in gray. Along each branch, the top number shows the number of genes concordant with the species tree at the associated node, and the bottom number represents the number of genes discordant with the species tree for the clade of interest. The color-coding of species names indicates tribe/subfamily membership and is consistent with **Figures 1–3**. As in **Figure 1**, select nodes are labeled with letters or abbreviations to enable easy reference in the text. Amygdaloideae tribes are labeled to the right of the species names.

TABLE 2 | The results of the *phyckle* analysis investigating gene tree support for alternative topologies regarding the phylogenetic placement of the tribe Spiraeeae.

Conflict	Topology	Bipartition	Number genes	Sum lnL difference
Phylogenetic position of Spiraeeae	nuclear	(*Aruncus, Holodiscus, Petrophytum, Cercocarpus, Dryas, Potentilla, Rubus, Ziziphus, Physocarpus*) | (all other taxa)	245	2503.2
	chloroplast	(*Aruncus, Holodiscus, Petrophytum, Sorbus, Vauquelinia, Amelanchier, Cydonia, Gillenia, Kageneckia, Malus, Rhaphiolepis*) | (all other taxa)	203	3115.8

For the chloroplast and nuclear topologies, the conflicting bipartitions, the number of genes supporting each relationship and the sum of log-likelihood differences for genes supporting each bipartition are shown.

tree than congruent with the species tree generally reflected nodes which disagree between the nuclear and chloroplast phylogenies, even though the data used for this analysis were nuclear gene trees and the nuclear species tree. The nodes with high conflict included deep nodes such as those displaying uncertainty regarding subfamilial relationships (node V; **Figure 4**) and the one reflecting the uncertainty of the position of the Spiraeeae tribe relative to the other tribes of the Amygdaloideae (node X; **Figure 4**). Moreover, the sister relationship between Amygdaleae + Lyonothamneae and a clade comprised of five other tribes (Sorbarieae, Kerrieae, Exochordeae, Gillenieae, and Maleae) showed high gene tree/species tree conflict (node Y; **Figure 4**). One other relatively deep node, representing the clade Kerrieae + Exochordeae + Sorbarieae (node KESo; **Figure 4**) exhibited high gene tree/species tree conflict, with over twice as many gene trees discordant as concordant. There were also several nodes with high degrees of discord within the Maleae, but investigating these shallower relationships is beyond the scope of this study, and we focused our taxon sampling with the goal of investigating deeper relationships in the tree as opposed to investigating documented discordance within the Maleae. When no bootstrap cutoff was used to consider whether gene trees were informative for a given node, the results were qualitatively similar (**Supplementary Figure 5**), so we focused on reporting the proportions of gene trees using the 50% bootstrap threshold (**Figure 4**). Based on the results of the SNaQ analysis, we used the *phyckle* "alternative relationship test" to further investigate support for the placement of the Spiraeeae using nuclear genes. We found that over 45% of nuclear genes (203 out of 448) support the chloroplast topology over the nuclear topology regarding the placement of Spiraeeae (**Table 2**). Moreover, the sum of log-likelihood differences across all genes indicated greater gene

tree support for the chloroplast topology than the nuclear topology (**Table 2**).

Multiply-Labeled Tree Analysis

The GRAMPA analysis revealed that an allopolyploid event likely occurred in the clade that resulted in Gillenieae + Maleae. The most parsimonious tree (score = 14,733) was a MUL-tree with multiple tips of all taxa within the Gillenieae + Maleae, with one clade sister to Exochordeae + Kerrieae + Sorbarieae, and one clade sister to the Spiraeeae (**Figure 5**). This MUL-tree was more parsimonious than the singly labeled tree (score = 14,777), which is considered evidence of allopolyploidy. The result that the Spiraeeae are one parental participant in an allopolyploidy event was consistent with the SNaQ network results. One difference between the most parsimonious GRAMPA MUL-tree and the optimal SNaQ network was that the GRAMPA tree shows Exochordeae + Kerrieae + Sorbarieae as sister to the Gillenieae + Maleae, whereas in the SNaQ network, an ancestor of the Sorbarieae was one half of the hybridization edge (**Figures 3**, **5**).

Hybridization Analyses

Hybrid Detector analyses confirmed aspects of both the SNaQ and GRAMPA results (**Table 3**). While using the Neillieae as an outgroup, the HyDe analysis inferred that the clade Maleae + Gillenieae was a hybrid with parents Spiraeeae and Sorbarieae, confirming the phylogenetic network result, and rejecting the possibility that either parental lineage (i.e., Spiraeeae or Sorbarieae) could be the hybrid lineage in this case (**Table 3**). The γ-value from the test that showed Maleae + Gillenieae as a hybrid lineage was 0.262 (**Table 3**). A similar analysis to test the relationship found using GRAMPA recovered support for Maleae + Gillenieae as a hybrid lineage with parents Spiraeeae and Kerrieae + Exochordeae + Sorbarieae (**Table 3**). Here the γ-value when Maleae + Gillenieae were a hybrid lineage was 0.526 (**Table 3**).

DISCUSSION

The important role of hybridization and genome doubling in generating plant diversity is becoming apparent (Soltis and Soltis, 2009; Folk et al., 2018). However, there are few well-supported examples of large, successful groups such as Maleae originating via wide hybridization and/or allopolyploidy from the ancestor of a small lineage (i.e., *Gillenia* Moench) (Evans and Campbell, 2002). Based on several complementary analyses—the comparison of nuclear and chloroplast phylogenies, phylogenetic network analyses, and allopolyploidy analyses using MUL-trees—we test the competing wide hybridization and spiraeoid hypotheses, and investigate the role of genome doubling, to explain the origin of the apple tribe. Here, we present multiple lines of evidence indicating that an ancestor of the Spiraeeae was likely the maternal participant in an ancient hybridization event and an ancestor of the clade Sorbarieae + Exochordeae + Kerrieae was likely the paternal participant, although there was some minor variation in analyses regarding the identity of the paternal parent (**Figures 3**, **5**). This hybridization event likely explains the origin of the clade Gillenieae + Maleae (**Figure 3**). Our results indicate that aspects of both existing hypotheses explaining the origin of the apple tribe are correct, but also aspects of each were incorrect. Our results also indicate that nuclear gene tree-species tree conflict and/or cytonuclear conflict are pervasive at several nodes in the Amygdaloideae. This suggests that beyond the hybrid origin of the apple clade, other lineages in the Amygdaloideae have reticulate evolutionary histories characterized by hybridization and/or allopolyploidy. Below, we discuss the details of our results and their implications on the origin of the apple tribe, as well as the possible explanations for high conflict nodes elsewhere in the subfamily Amygdaloideae.

The Ancient Hybrid Origin of Maleae-Gillenieae and Subsequent Genome Doubling in Maleae

Our results suggest the ancestor of the Maleae + Gillenieae originated via hybridization between distantly related tribes in the Amygdaloideae (i.e., the wide hybridization hypothesis, which states that the Maleae are the result of an ancient hybridization event between progenitors from other tribes in the subfamily Amygdaloideae) (**Figure 3**). Specifically, there was a hybridization event between an ancestor of the tribe Spiraeeae (x = 9) and an ancestor of Sorbarieae (x = 9) + Exochordeae (x = 8) + Kerrieae (x = 9), which gave rise to the clade comprised of Gillenieae (x = 9) + Maleae (x = 17) (**Figure 3**). This result is largely congruent with the wide hybridization hypothesis, except that we found that the clade Gillenieae + Maleae was the result of a wide hybridization event, as opposed to just the Maleae (**Figure 3**). Our results also partially support the spiraeoid hypothesis (i.e., the 17 chromosomes found in Maleae arose via the genome doubling of an x = 9 ancestor to x = 18, and subsequent aneuploidy resulting in x = 17), specifically regarding the role of whole genome duplication in the origin of the Maleae (**Figure 5**). The ancestors involved in the hybridization event leading to Gillenieae + Maleae had base chromosome numbers of x = 8 or 9, so there may have been genome doubling, possibly followed by aneuploidy if two x = 9 taxa were involved, to result in the x = 17 observed in the Maleae (**Figure 5**). Regardless of the ancestral chromosome number (x = 8 vs. x = 9), the genome doubling aspect of the spiraeoid hypothesis is supported by our results. However, given that our network analysis found that the clade Gillenieae + Maleae was the result of a hybridization event, and the base chromosome number of Gillenieae is x = 9, a genome doubling event preceding Gillenieae + Maleae can readily explain the x = 17 observed in the Maleae but not x = 9 in Gillenieae (**Figure 3**). The Gillenieae lineage may have undergone diploidization following an allopolyploidy event whereas the Maleae did not. Interpretation of the GRAMPA analysis favors this explanation because the most parsimonious allopolyploidy event precedes Gillenieae + Maleae, as opposed to only Maleae. Alternatively, perhaps there was a second genome doubling event of Maleae after an initial hybridization event leading to Gillenieae + Maleae. We favor the latter explanation, which is consistent with our SNaQ analysis and with results

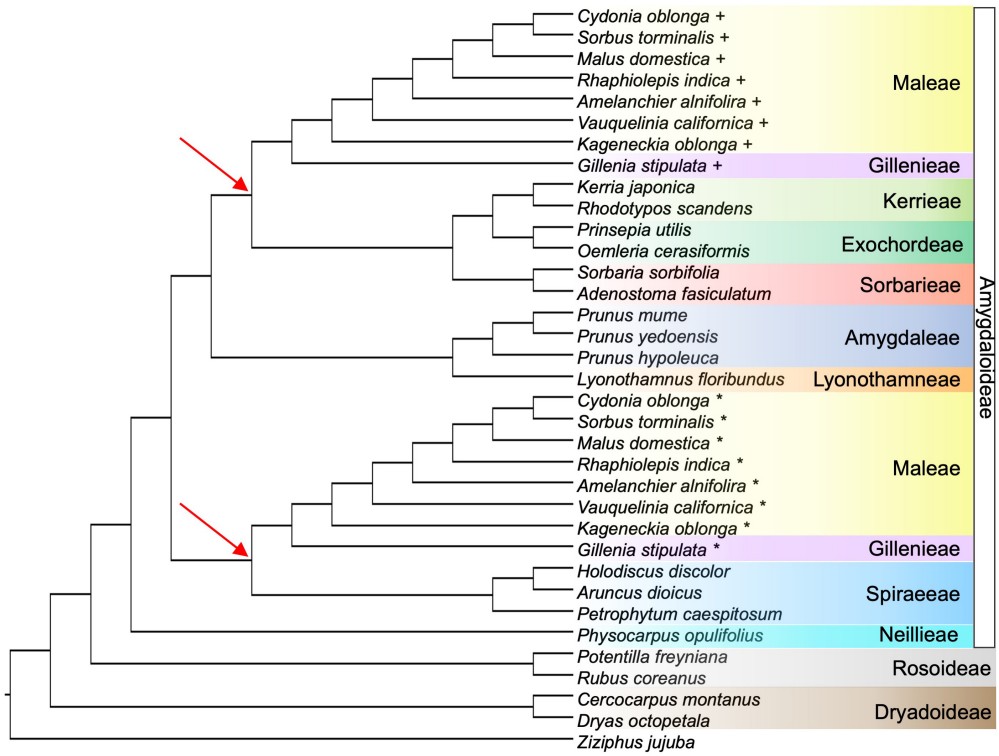


FIGURE 5 | The most parsimonious tree from the GRAMPA analysis, which is a multi-labeled (MUL) tree indicating two tips for all species in Maleae + Gillenieae. For each species with multiple labels, the first tip is indicated by a plus sign and the second tip is shown using an asterisk. One Gillenieae + Maleae clade is sister to Kerrieae + Exochordeae + Sorbarieae, and the other Gillenieae + Maleae clade is sister to Spiraeeae. The red arrows highlight the nodes defining different lineages sister to the multi-labeled taxa in Maleae + Gillenieae. The color-coding of species shows tribe/subfamily and is consistent with all previous figures.

TABLE 3 | The two hybridization hypotheses tested using Hybrid Detector.

	Parent 1	Hybrid	Parent 2	γ-value	Z-score	*P*-value
Hybrid relationship inferred by SNaQ	Sorbarieae	Maleae-Gillenieae	Spiraeeae	0.262	14.321	0.000
	Sorbarieae	Spiraeeae	Maleae-Gillenieae	–1.233	–99999.9	1.000
	Spiraeeae	Sorbarieae	Maleae-Gillenieae	0.608	–25.934	1.000
Hybrid relationship inferred by GRAMPA	Spiraeeae	Maleae-Gillenieae	Kerrieae-Exochordeae-Sorbarieae	0.526	21.569	0.000
	Spiraeeae	Kerrieae-Exochordeae-Sorbarieae	Maleae-Gillenieae	0.911	–2.349	0.991
	Maleae-Gillenieae	Spiraeeae	Kerrieae-Exochordeae-Sorbarieae	–0.122	–2.118	0.983

The first HyDe analysis (top) found support for the hypothesis of Maleae-Gillenieae as a hybrid taxon resulting from parents Sorbarieae and Spiraeeae, which is consistent with the SNaQ result. The second HyDe analysis (bottom) inferred that Maleae-Gillenieae was a hybrid of parents Spiraeeae and Kerrieae-Exochordeae-Sorbarieae, which corresponds to the GRAMPA results.

from Xiang et al. (2017), who noted nodes in the Maleae with evidence of whole genome duplications (WGDs) after the divergence of the Maleae from the ancestor of Gillenieae + Maleae (see Xiang et al., 2017; **Figure 5**). The annotated genome assembly of another Gillenieae species, *Gillenia trifoliata*, revealed that many syntenic blocks in *Gillenia trifoliata* mapped to two locations in *Malus domestica*, as would be expected with a history of genome doubling (Ireland et al., 2021). Moreover, the same syntenic blocks correspond to single orthologous regions in other Rosaceae species [*Rubus occidentalis* (raspberry) of Rosoideae and *Prunus persica* (peach)] of Amygdaloideae, suggesting that it is unlikely that the Gillenieae underwent a WGD and subsequent diploidization—a simpler explanation is that the WGD occurred after an initial hybridization leading to Gillenieae + Maleae.

Across the plant tree of life, diversification via genome duplication is relatively common. It is becoming increasingly clear that following WGD events, the genomes of organisms are particularly malleable and that genomic rearrangements may spur key functional innovations. Genome evolution associated with WGDs has often been studied in crop species, many of which are polyploid. For example, controlled crosses of early generation allopolyploid wheat revealed that aneuploidy is common following WGDs (Zhang et al., 2013). However,

there are examples of variation in genome size and organization after WGD events in non-model systems. In the neopolyploid *Tragopogon* L., massive chromosomal variation followed an allopolyploidy event (Chester et al., 2012), including aneuploidy in 69% of cases. Within a single genus of ca. 250 species (clover; genus *Trifolium* L.), there have been many deviations from the ancestral chromosome state ($2n$ = 16), including at least 22 instances of polyploidy and 19 occurrences of aneuploidy (Ellison et al., 2006). The diversification of the Gillenieae-Maleae clade may represent another example of lineages that diversified following chromosomal rearragnements via allopolyploidy and aneuploidy.

Many previous studies have hypothesized an allopolyploid origin of the apple tribe (Sax, 1931, 1932, 1933; Stebbins, 1950; Chevreau et al., 1985; Weeden and Lamb, 1987; Robertson et al., 1991; Evans and Campbell, 2002; Vamosi and Dickinson, 2006; Potter et al., 2007). The wide-hybridization hypothesis, favored until 2002, considered many lineages as possible participants in hybridization and/or allopolyploidy, but strong evidence for any particular lineage was lacking. The spiraeoid hypothesis was supported by one duplicated nuclear gene (*GBSSI-1* and *GBSSI-2*; Evans and Campbell, 2002), inferring that both parental participants in allopolyploidy were ancestors of the Gillenieae lineage. Our network analyses (**Figure 3**) indicate that a hybridization event between an ancestor of the Spiraeeae and Sorbarieae leading to Gillenieae + Maleae, whereas allopolyploid analyses (**Figure 5**) indicate that an ancestor of the Spiraeeae and a common ancestor of Sorbarieae + Exochordeae + Kerrieae were likely the parental participants in allopolyploidy. Both of these scenarios were confirmed as possible hybridization events using separate analyses (i.e., HyDe; **Table 3**). Given the low support for the KESo and Z nodes (quartet support scores = 50.44 and 39.70, respectively, and high degrees of gene tree conflict; **Figure 4**), perhaps the topological uncertainty in the nuclear phylogeny is causing the discrepancy between the SNaQ and GRAMPA analyses (**Figures 3**, **5**). When considering the bifurcating nuclear and plastome topologies (**Figure 2**), and considering the proportions of nuclear gene trees that support the nuclear vs. chloroplast topologies (**Table 2**), it becomes evident that the Spiraeeae ancestor was most likely the maternal donor to a hybridization or allopolyploid event because Spiraeeae is sister to Gillenieae + Maleae in the plastome tree, and that the ancestor of Sorbarieae + Exochordeae + Kerrieae was the paternal participant because this relationship is more similar to the nuclear tree than the plastome tree.

Discordance/Reticulation Throughout the Amygdaloideae

There are multiple nodes with pervasive conflict, both among the subfamilies of the Rosaceae and within the Amygdaloideae. These include nodes V (Rosoideae—Amygdaloideae sister), Y (Lyonothamneae + Amygdaleae sister to clade defined by node Z), L (Lyonothamneae sister to Amygdaleae), Z (Kerrieae + Exochordeae + Sorbarieae sister to Gillenieae +Maleae), and KESo (Sorbarieae sister to Kerrieae + Maleae) (**Figure 4**). Xiang et al. (2017) produced six distinct Rosaceae phylogenies based on data filtering (between 113 and 882 genes included) and tree-inference method (concatenation with ML inference in RAxML vs. a coalescent species tree approach implemented in ASTRAL). They defined nodes as highly supported (100% bootstrap support in all trees), moderately supported (90% bootstrap support in at least five trees, and 85% support in all six trees), poorly supported (80% bootstrap support in three or more trees and 40% support in all six trees), and unresolved (not meeting the above criteria). Multiple nodes with pervasive conflict according to our *phyparts* analysis were considered highly supported (Y, V, L) or moderately supported (KESo) in Xiang et al. (2017). Only one key node with high gene tree conflict (Z) was considered poorly supported in Xiang et al. (2017), and no nodes with pervasive conflict identified in our analyses were listed as unresolved. None of the above nodes were consistent with the plastome tree, either from Zhang et al. (2017) or from our analyses. Clearly, there is substantial conflict among nuclear gene trees within the Amygdaloideae, in addition to the documented cytonuclear discord. Histories of reticulate evolution appear common in this group, beyond the allopolyploid origin of the apple tribe.

We cannot be certain of the cause of conflict in many of the nodes in the Amygdaloideae. Potential biological explanations for gene tree discord may include incomplete lineage sorting (ILS) or hybridization. Processes such as ILS may also lead to gene tree-species tree conflict in the absence of hybridization. However, there is strong evidence for the hybrid origin of the Gillenieae + Maleae. SNaQ is robust to ILS in that it can incorporate uncertainty in user-estimated gene trees and handle gene tree discordance caused by ILS (Solís-Lemus and Ané, 2016). The comparison of pseudodeviance network scores in SNaQ between the ASTRAL tree, which accommodates ILS, and the h_{max} = 1 network, which can accommodate ILS and hybrid edges, clearly favors the h_{max} = 1 network. The *phyckle* analysis of the node defining the position of the Spiraeeae (i.e., node X, **Figure 4** and **Table 2**) on its own can identify the proportion of nuclear genes that support species tree or alternative relationships, but does not explicitly identify sources of conflict. However, the nearly equal distribution of nuclear genes that support the nuclear topology and the chloroplast topology, when considered alongside the other analyses (e.g., SNaQ, GRAMPA, and HyDe), add evidence that a history of hybridization via allopolyploidy shaped evolutionary histories of the sampled genes. While we do not have specific expectations for the proportion of gene trees that may conflict with the species tree solely due to ILS, that so many gene trees support the alternative chloroplast topology, as opposed to a distribution of different topologies induced by ILS, provides more evidence for an instance of hybridization. That over 45% of nuclear genes (203 out of 448) support the chloroplast topology over the nuclear topology with regard to the placement of Spiraeeae is another piece of evidence that the maternal participant in allopolyploidy leading to the apple tribe was an ancestor of the Spiraeeae. The large number of nuclear genes that favor the chloroplast topology may in part explain past uncertainty in phylogenetic studies investigating the Rosaceae or its subfamilies.

The tribe Lyonothamneae is represented by a monotypic genus, *Lyonothamnus* A.Gray. The position of this tribe varies greatly between the plastome phylogeny (Lyonothamneae sister to all other tribes in the Amygdaloideae) and nuclear phylogeny (Lyonothamneae sister to Amygdaleae). Furthermore, there is substantial nuclear gene tree conflict at this node (L; **Figure 4**). The SNaQ network with h_{max} = 2 showed that the second hybridization edge was between *Prunus hypoleuca* of the *Maddenia* group and the lineage ancestral to Lyonothamneae + Amygdaleae (**Supplementary Figure 4**). Essentially, this hybridization edge means that *Prunus hypoleuca* of the *Maddenia* group (Wen and Shi, 2012) is 87.9% sister to the other *Prunus* species, and 12.1% sister to the ancestor of Lyonothamneae + Amygdaleae. The interpretation of this hybridization edge is less straightforward than the h_{max} = 1 edge. However, there is evidence from previous studies that a WGD occurred near the base of the Amygdaleae (Xiang et al., 2017), and other studies have hypothesized that ancient hybridization and/or allopolyploidy were involved in the diversification of *Prunus* (Chin et al., 2014; Zhao et al., 2016, 2018; Hodel et al., 2021), the sole accepted genus in the Amygdaleae. Future studies with denser taxon-sampling in the Amygdaleae and hundreds of nuclear loci combined with chloroplast data are needed to investigate the evolutionary history of the Lyonothamneae + Amygdaleae.

Although assessing discordance within the Maleae is not a focus of this paper, we note that in the h_{max} = 4 and h_{max} = 5 SNaQ networks (**Supplementary Figure 4**), there are hybridization edges that indicate possible hybridization within the Maleae. The h_{max} = 4 hybrid edge shows *Kageneckia* 96.2% sister to all other Maleae but also 3.8% sister to Maleae + Gillenieae. In the h_{max} = 5 network, the hybrid edge indicates the ancestor of subtribe Malinae (pome-bearing Maleae, i.e., Maleae excluding *Kageneckia* and *Vauquelinia*) is 97.7% sister to *Vauquelinia* and also 2.3% sister to Maleae + Gillenieae. Taken in isolation, these hybrid edges mean little, especially given the discrepancy between the γ values of the major and minor hybridization edges. However, when considered in concert with previously documented discord within the Maleae, the conflict documented via *phyparts* at multiple nodes in the Maleae (**Figure 4**), as well as evidence of genome doubling at multiple nodes within the Maleae (Xiang et al., 2017), the h_{max} = 4 and h_{max} = 5 SNaQ results point to hybridization, especially introgression, as a possible mechanism explaining phylogenomic discord within the Maleae. Further targeted investigations are needed to address discordance within the Maleae.

The well-documented discordance between chloroplast and nuclear phylogenies in the Amygdaloideae could also be explained by chloroplast capture. This phenomenon occurs when native cytoplasm is replaced by foreign cytoplasm via hybridization followed by repeated backcrossing (Rieseberg and Soltis, 1991). In closely related species that are sexually compatible, chloroplast capture can be pervasive and lead to cytonuclear discordance. In the Amygdaloideae, there have been several instances of chloroplast capture documented. In the Amygdaleae tribe, cytonuclear discord was attributed to chloroplast capture in several *Prunus* species, including North American plums (Rohrer et al., 2008) and East Asian cherries (Cho et al., 2014). Within the Maleae, there is also evidence of chloroplast capture as a mechanism causing cytonuclear discord. Strong discordance between nuclear and plastid phylogenies regarding the placement of the Maleae genera *Malacomeles* (Decne.) Decne. and *Peraphyllum* Nutt. supports ancient chloroplast capture events in SW North America (Liu et al., 2020a). Because chloroplast capture involves hybridization followed by recurrent backcrossing, it occurs more frequently at shallower systematic scales among sexually compatible species. Accordingly, chloroplast capture could explain the SNaQ hybridization edges detected within the Maleae and Amygdaleae at higher values of h_{max} (**Supplementary Figure 4**). Additionally, although we did not detect cytonuclear discord within the Maleae in our sampling, histories consistent with chloroplast capture may explain the pervasive gene tree conflict at nodes within the Maleae (**Figure 4**).

Synthesizing Multiple Nuclear Genes and Chloroplast Data Resolves Cases of Reticulation

Previous molecular studies of the Rosaceae typically used either nuclear (e.g., Evans and Campbell, 2002) or chloroplast data (Potter et al., 2002). The single-nuclear gene *GBSSI* phylogeny by Evans and Campbell (2002) could not resolve the position of the Spiraeeae, and the branching order of the Spiraeeae, Sorbarieae, Exochordeae, Amygdaleae, and Dryadeae was a polytomy. Potter et al. (2002) used two chloroplast genes and recovered a phylogeny that placed the Spiraeeae + Sorbarieae sister to the Gillenieae + Maleae + Amygdaleae. However, there was poor bootstrap support (i.e., < 75%) for all these relationships except Gillenieae + Maleae. Potter et al.'s (2002) chloroplast phylogeny also found that Lyonothamneae were sister to all other Amygdaloid tribes with 100% bootstrap support. One study that used data from both nuclear and chloroplast genomes is Potter et al. (2007), with six nuclear and four chloroplast loci to create a consensus phylogeny, inferred that the Spiraeeae were sister to the Gillenieae + Maleae (i.e., the dominant chloroplast topology from Zhang et al. (2017) and the present study), albeit with low support (44% bootstrap and 57% Bayesian clade credibility). Potter et al. (2007) excluded two nuclear loci from their analyses due to results "inconsistent in some ways with the majority of other data." Two of the anomalous results caused by the two excluded nuclear genes they report are inconsistent placement of the Spiraeeae and the lack of a sister relationship between Lyonothamneae and the rest of the Amygdaloideae (referred to as Spiraeoideae in Potter et al., 2007).

Subsequent results, from Xiang et al. (2017) and Zhang et al. (2017) and the present study, contextualize and explain the results from earlier molecular studies. The position of Lyonothamneae is clearly quite different in the chloroplast and nuclear genomes, and this is reflected throughout the literature; studies with only chloroplast data repeatedly find Lyonothamneae sister to the rest of Amygdaloideae, typically with strong support. In contrast, this relationship

is never found in studies using only nuclear data. The position of the Spiraeeae has been variable in studies from the literature, but it is now becoming clear that much of the uncertainty with regard to its placement is due to a history of WGDs in the Amygdaloideae. Specifically, in this paper we characterize one instance of allopolyploidy, in which an ancestor of the Spiraeeae was likely the maternal participant in allopolyploidization. Given the distinct chloroplast and nuclear topologies regarding the placement of Spiraeeae, and the fact that nearly half of nuclear genes sampled in this study favor the chloroplast topology, it is unsurprising that earlier molecular studies using fewer than 10 markers were unable to confidently resolve the position of Spiraeeae. Although we used a set of complementary analyses to resolve the origin of the apple tribe, there is clearly more phylogenetic uncertainty due to reticulate evolutionary histories in the Amygdaloideae. The different positions of the Lyonothamneae in the nuclear and chloroplast phylogenies, coupled with the SNaQ network results and previous evidence of WGD events leading to and within the Amygdaleae, indicate that future targeted efforts should be focused on resolving the evolutionary history of the Lyonothamneae and Amygdaleae.

CONCLUSION AND PROSPECTS

Over the past several decades, systematists have embraced the need for incorporating genealogical information from nuclear genes to obtain robust estimates of phylogeny. Chloroplast data were favored for many years due to their high copy number which translated to easy generation of homologous loci for many individuals and/or species (Thode et al., 2020; Wang et al., 2020; Welker et al., 2020). Those studies fell out of favor due to the limited information regarding ancestry given their typical uniparental inheritance. It is now becoming clear that reticulation is prevalent at many phylogenetic scales due to hybridization or other processes (Lee-Yaw et al., 2018). In cases of pervasive reticulation, nuclear AND chloroplast data are now necessary complements to one another if researchers hope to resolve reticulate complexes. Our study highlights how synthesizing results from existing studies cannot only reconcile differences from two recent studies, but also answer century old questions that have been continually debated in the literature. Our study also highlights the need to revisit and reconsider phylogenetic relationships, even when they have been found to be highly supported using metrics such as bootstrapping. In a number of recent studies (e.g., Soltis et al., 2004; Prasanna et al., 2020; Walker et al., 2021), careful analyses of conflict have revealed that we should not be overly confident in apparently resolved relationships. In conclusion, our results from multiple lines of evidence confirmed the hybrid origin of the Maleae + Gillenieae clade and supported the polyploidy-aneuploidy-origin aspect of the hypothesis of Maleae ($x = 17$ or 15) originating from the tribe Gillenieae ($x = 9$) as proposed by Evans and Campbell (2002). Future research may provide a complete picture of the role of hybridization in the early diversification of Maleae, especially regarding the formation of the chromosome number of 15 in *Vauquelinia* and the evolutionary mechanisms leading from dry fruits (capsules) to fleshy fruits (pomes).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

RH, EZ, and JW conceptualized the project. RH and B-BL obtained and analyzed the data. RH led the writing of the manuscript. All authors edited drafts and approved the final version of the manuscript.

FUNDING

This work was supported by a Smithsonian Institution Peter Buck Fellowship to RH.

ACKNOWLEDGMENTS

We thank Greg Stull, two reviewers, and the associate editor for many helpful comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.820997/full#supplementary-material

REFERENCES

Baudry, J.-P., Maugis, C., and Michel, B. (2011). Slope heuristics: overview and implementation. *Stat. Comput.* 22, 455–470. doi: 10.1007/s11222-011-9236-1

Blischak, P. D., Chifman, J., Wolfe, A. D., and Kubatko, L. S. (2018). HyDe: a python package for genome-scale hybridization detection. *Syst. Biol.* 67, 821–829. doi: 10.1093/sysbio/syy023

Brown, J. W., Walker, J. F., and Smith, S. A. (2017). Phyx: phylogenetic tools for unix. *Bioinformatics* 33, 1886–1888. doi: 10.1093/bioinformatics/btx063

Bruun-Lund, S., Clement, W. L., Kjellberg, F., and Rønsted, N. (2017). First plastid phylogenomic study reveals potential cyto-nuclear discordance in the evolutionary history of *Ficus* L. (*Moraceae*). *Mol. Phylogenet. Evol.* 109, 93–104. doi: 10.1016/j.ympev.2016.12.031

Campbell, C. S., Donoghue, M. J., Baldwin, B. G., and Wojciechowski, M. F. (1995). Phylogenetic relationships in *Maloideae* (Rosaceae): evidence from sequences of the internal transcribed spacers of nuclear ribosomal DNA and its congruence with morphology. *Am. J. Bot.* 82, 903–918. doi: 10.1002/j.1537-2197.1995.tb15707.x

Chester, M., Gallagher, J. P., Symonds, V. V., Da Silva, A. V. C., Mavrodiev, E. V., Leitch, A. R., et al. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc. Natl. Acad. Sci. U.S.A.* 109, 1176–1181. doi: 10.1073/pnas.1112041109

Chevreau, E., Lespinasse, Y., and Gallet, M. (1985). Inheritance of pollen enzymes and polyploid origin of apple (*Malus x domestica* Borkh.). *Theor. Appl. Genet.* 71, 268–277. doi: 10.1007/BF00252066

Chin, S.-W., Shaw, J., Haberle, R., Wen, J., and Potter, D. (2014). Diversification of almonds, peaches, plums and cherries – molecular systematics and biogeographic history of *Prunus* (Rosaceae). *Mol. Phylogenet. Evol.* 76, 34–48. doi: 10.1016/j.ympev.2014.02.024

Cho, M. S., Kim, C. S., Kim, S. H., Kim, T. O., Heo, K. I., Jun, J., et al. (2014). Molecular and morphological data reveal hybrid origin of wild *Prunus yedoensis* (Rosaceae) from Jeju Island, Korea: implications for the origin of the flowering cherry. *Am. J. Bot.* 101, 1976–1986. doi: 10.3732/ajb.1400318

Darlington, C. D., and Moffett, A. A. (1930). Primary and secondary chromosome balance in *Pyrus*. *J. Genet.* 22, 129–151. doi: 10.1007/bf02983843

Ellison, N. W., Liston, A., Steiner, J. J., Williams, W. M., and Taylor, N. L. (2006). Molecular phylogenetics of the clover genus (*Trifolium-Leguminosae*). *Mol. Phylogenet. Evol.* 39, 688–705. doi: 10.1016/j.ympev.2006.01.004

Evans, R. C., and Campbell, C. S. (2002). The origin of the apple subfamily (Maloideae; Rosaceae) is clarified by DNA sequence data from duplicated *GBSSI* genes. *Am. J. Bot.* 89, 1478–1484. doi: 10.3732/ajb.89.9.1478

Folk, R. A., Soltis, P. S., Soltis, D. E., and Guralnick, R. (2018). New prospects in the detection and comparative analysis of hybridization in the tree of life. *Am. J. Bot.* 105, 364–375. doi: 10.1002/ajb2.1018

Gladkova, V. N. (1972). On the origin of subfamily Maloideae. *Bot. Zhur.* 57, 42–49.

Goldblatt, P. (1976). Cytotaxonomic studies in the tribe Quillajeae (Rosaceae). *Ann. Miss. Bot. Gard.* 63, 200–206. doi: 10.2307/2395226

Goodman, M., Czelusniak, J., William Moore, G., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Biol.* 28, 132–163. doi: 10.2307/2412519

Hodel, R. G. J., Zimmer, E., and Wen, J. (2021). A phylogenomic approach resolves the backbone of *Prunus* (Rosaceae) and identifies signals of hybridization and allopolyploidy. *Mol. Phylogenet. Evol.* 160:107118. doi: 10.1016/j.ympev.2021.107118

Huang, D. I, Hefer, C. A., Kolosova, N., Douglas, C. J., and Cronk, Q. C. B. (2014). Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). *New Phytol.* 204, 693–703. doi: 10.1111/nph.12956

Ireland, H. S., Wu, C., Deng, C. H., Hilario, E., Saei, A., Erasmuson, S., et al. (2021). The *Gillenia trifoliata* genome reveals dynamics correlated with growth and reproduction in Rosaceae. *Hortic. Res.* 8:233. doi: 10.1038/s41438-021-00662-4

Katoh, K., and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Larget, B. R., Kotha, S. K., Dewey, C. N., and Ané, C. (2010). Bucky: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* 26, 2910–2911. doi: 10.1093/bioinformatics/btq539

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., and Rieseberg, L. H. (2018). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). *New Phytol.* 221, 515–526. doi: 10.1111/nph.15386

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

Liu, B.-B., Campbell, C. S., Hong, D. Y., and Wen, J. (2020a). Phylogenetic relationships and chloroplast capture in the *Amelanchier-Malacomeles-Peraphyllum* clade (Maleae, Rosaceae): evidence from chloroplast genome and nuclear ribosomal DNA data using genome skimming. *Mol. Phylogenet. Evol.* 147:106784. doi: 10.1016/j.ympev.2020.106784

Liu, B.-B., Hong, D.-Y., Zhou, S. L., Xu, C., Dong, W. P., Johnson, G., et al. (2019). Phylogenomic analyses of the *Photinia* complex support the recognition of a new genus *Phippsiomeles* and the resurrection of a redefined *Stranvaesia* in Maleae (Rosaceae). *J. Syst. Evol.* 57, 678–694. doi: 10.1111/jse.12542

Liu, B.-B., Liu, G.-N., Hong, D.-Y., and Wen, J. (2020b). *Eriobotrya* belongs to *Rhaphiolepis* (Maleae, Rosaceae): evidence from chloroplast genome and nuclear ribosomal DNA data. *Front. Plant Sci.* 10:1731. doi: 10.3389/fpls.2019.01731

Liu, B.-B., Ren, C., Kwak, M., Hodel, R. G. J., Xu, C., He, J., et al. (2021). Phylogenomic analyses in the apple genus *Malus* s.l. reveal widespread hybridization and allopolyploidy driving the diversifications, with insights into the complex biogeographic history in the Northern Hemisphere. *bioRxiv* [Preprint] doi: 10.1101/2021.10.12.464085

Mirarab, S., Reaz, R., and Bayzid, M. S. (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30, i541–i548.

Nebel, B. (1929). Zur cytologie von malus und vitis. *Die Gart. Bauwiss.* 1, 549–592.

Page, R. D. M. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* 43, 58–77. doi: 10.1093/sysbio/43.1.58

Potter, D., Eriksson, T., Evans, R. C., Oh, S., Smedmark, J. E. E., Morgan, D. R., et al. (2007). Phylogeny and classification of Rosaceae. *Plant Syst. Evol.* 266, 5–43.

Potter, D., Gao, F., Bortiri, P. E., Oh, S.-H., and Baggett, S. (2002). Phylogenetic relationships in Rosaceae inferred from chloroplast *matK* and *trnL-trnF* nucleotide sequence data. *Plant Syst. Evol.* 231, 77–89.

Prasanna, A. N., Gerber, D., Kijpornyongpan, T., Aime, M. C., Doyle, V. P., and Nagy, L. G. (2020). Model choice, missing data, and taxon sampling impact phylogenomic inference of deep basidiomycota relationships. *Syst. Biol.* 69, 17–37. doi: 10.1093/sysbio/syz029

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. doi: 10.1111/j.2041-210x.2011.00169.x

Rieseberg, L. H., and Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trends Plants* 5, 65–84.

Robertson, K. R., Phipps, J. B., Rohrer, J. R., and Smith, P. G. (1991). A synopsis of genera in Maloideae (Rosaceae). *Syst. Bot.* 16:376. doi: 10.2307/2419287

Rohrer, J. R., O'brien, M. A., and Anderson, J. A. (2008). Phylogenetic analysis of North American plums (*Prunus* sect. *Prunocerasus*: Rosaceae) based on nuclear *Leafy* and *s6pdh* sequences. *J. Bot. Res. Inst. Tx* 2, 401–414.

Roycroft, E. J., Moussalli, A., and Rowe, K. C. (2020). Phylogenomics uncovers confidence and conflict in the rapid radiation of Australo-Papuan rodents. *Syst. Biol.* 69, 431–444. doi: 10.1093/sysbio/syz044

Sax, K. (1931). The origin and relationships of the Pomoideae. *J. Arnold Arbor.* 12, 3–22.

Sax, K. (1932). Chromosome relationships in Pomoideae. *J. Arnold Arbor.* 13, 363–367.

Sax, K. (1933). The origin of the Pomoideae. *Proc. Am. Soc. Hortic. Sci.* 30, 147–150. doi: 10.1046/j.1365-2672.1997.00377.x

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* 15:150. doi: 10.1186/s12862-015-0423-0

Smith, S. A., Walker-Hale, N., Walker, J. F., and Brown, J. W. (2020). Phylogenetic conflicts, combinability, and deep phylogenomics in plants. *Syst. Biol.* 69, 579–592. doi: 10.1093/sysbio/syz078

Solís-Lemus, C., and Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 12:e1005896. doi: 10.1371/journal.pgen.1005896

Solís-Lemus, C., Bastide, P., and Ané, C. (2017). PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.* 34, 3292–3298. doi: 10.1093/molbev/msx235

Soltis, D. E., Albert, V. A., Savolainen, V., Hilu, K., Qiu, Y. L., Chase, M. W., et al. (2004). Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. *Trends Plant Sci.* 9, 477–483. doi: 10.1016/j.tplants.2004.08.008

Soltis, P. S., and Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.* 60, 561–588. doi: 10.1146/annurev.arplant.043008.092039

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

Stebbins, G. L. (1950). *Variation and Evolution in Plants*. New York, NY: Columbia University Press.

Struck, T. H. (2014). Trespex-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evol. Bioinformat.* 10, 51–67. doi: 10.4137/EBO.S14239

Thode, V. A., Lohmann, L. G., and Sanmartín, I. (2020). Evaluating character partitioning and molecular models in plastid phylogenomics at low taxonomic levels: a case study using *Amphilophium* (Bignonieae, Bignoniaceae). *J. Syst. Evol.* 58, 1071–1089.

Thomas, G. W. C., Ather, S. H., and Hahn, M. W. (2017). Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Syst. Biol.* 66, 1007–1018. doi: 10.1093/sysbio/syx044

Vamosi, J. C., and Dickinson, T. A. (2006). Polyploidy and diversification: a phylogenetic investigation in Rosaceae. *Int. J. Plant Sci.* 167, 349–358.

Vaughan, T. G. (2017). IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics* 33, 2392–2394. doi: 10.1093/bioinformatics/btx155

Walker, J. F., Smith, S. A., Hodel, R. G. J., and Moyroud, E. (2021). Concordance-based approaches for the inference of relationships and molecular rates with phylogenomic data sets. *Syst. Biol.* 7:syab052. doi: 10.1093/sysbio/syab052

Wang, H. X., Morales-Briones, D. F., Moore, M. J., Wen, J., and Wang, H. F. (2021). A phylogenomic perspective on gene tree conflict and character evolution in Caprifoliaceae using target enrichment data, with Zabelioideae recognized as a new subfamily. *J. Syst. Evol.* 59, 897–914.

Wang, Y.-B., Liu, B.-B., Nie, Z.-L., Chen, H.-F., Chen, F.-J., Figlar, R. B., et al. (2020). Major clades and a revised classification of *Magnolia* and *Magnoliaceae* based on whole plastid genome sequences via genome skimming. *J. Syst. Evol.* 58, 673–695. doi: 10.1111/jse.12588

Weeden, N., and Lamb, R. (1987). Genetics and linkage analysis of 19 isozyme loci in apple. *J. Am. Soc. Hortic. Sci.* 112, 865–872.

Welker, C. A. D., Mckain, M. R., Estep, M. C., Pasquet, R. S., Chipabika, G., Pallangyo, B., et al. (2020). Phylogenomics enables biogeographic analysis and a new subtribal classification of the Andropogoneae (Poaceae—Panicoideae). *J. Syst. Evol.* 58:10031030.

Wen, J., and Shi, W. (2012). Revision of the *Maddenia* clade of *Prunus* (Rosaceae). *PhytoKeys* 11, 39–59. doi: 10.3897/phytokeys.11.2825

Xiang, Y., Huang, C. H., Hu, Y., Wen, J., Li, S., Yi, T., et al. (2017). Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* 34, 262–281.

Xu, L.-L., Yu, R.-M., Lin, X.-R., Zhang, B.-W., Li, N., Lin, K., et al. (2021). Different rates of pollen and seed gene flow cause branch-length and geographic cytonuclear discordance within Asian butternuts. *New Phytol.* 232, 388–403. doi: 10.1111/nph.17564

Zhang, H., Bian, Y., Gou, X., Zhu, B., Xu, C., Qi, B., et al. (2013). Persistent whole-chromosome aneuploidy is generally associated with nascent allohexaploid wheat. *Proc. Natl. Acad. Sci. U.S. A.* 110, 3447–3452. doi: 10.1073/pnas.1300153110

Zhang, S. D., Jin, J. J., Chen, S. Y., Chase, M. W., Soltis, D. E., Li, H. T., et al. (2017). Diversification of Rosaceae since the late *Cretaceous* based on plastid phylogenomics. *New Phytol.* 214, 1355–1367. doi: 10.1111/nph.14461

Zhao, L., Jiang, X.-W., Zuo, Y., Liu, X.-L., Chin, S.-W., Haberle, R., et al. (2016). Multiple events of allopolyploidy in the evolution of the racemose lineages in *Prunus* (Rosaceae) based on integrated evidence from nuclear and plastid data. *PLoS One* 11:e0157123. doi: 10.1371/journal.pone.0157123

Zhao, L., Potter, D., Xu, Y., Liu, P. L., Johnson, G., Chang, Z. Y., et al. (2018). Phylogeny and spatio-temporal diversification of *Prunus* subgenus *Laurocerasus* section *Mesopygeum* (Rosaceae) in the Malesian region. *J. Syst. Evol.* 56, 637–651. doi: 10.1111/jse.12467

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 31 January 2022
doi: 10.3389/fpls.2021.824672

Organelle Phylogenomics and Extensive Conflicting Phylogenetic Signals in the Monocot Order Poales

Hong Wu[1,2], Jun-Bo Yang[1], Jing-Xia Liu[1], De-Zhu Li[1] and Peng-Fei Ma[1]**

[1] Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China,
[2] University of Chinese Academy of Sciences, Beijing, China

The Poales is one of the largest orders of flowering plants with significant economic and ecological values. Reconstructing the phylogeny of the Poales is important for understanding its evolutionary history that forms the basis for biological studies. However, due to sparse taxon sampling and limited molecular data, previous studies have resulted in a variety of contradictory topologies. In particular, there are three nodes surrounded by incongruence: the phylogenetic ambiguity near the root of the Poales tree, the sister family of Poaceae, and the delimitation of the xyrid clade. We conducted a comprehensive sampling and reconstructed the phylogenetic tree using plastid and mitochondrial genomic data from 91 to 66 taxa, respectively, representing all the 16 families of Poales. Our analyses support the finding of Bromeliaceae and Typhaceae as the earliest diverging groups within the Poales while having phylogenetic relationships with the polytomy. The clade of Ecdeiocoleaceae and Joinvilleaceae is recovered as the sister group of Poaceae. The three families, Mayacaceae, Eriocaulaceae, and Xyridaceae, of the xyrid assembly diverged successively along the backbone of the Poales phylogeny, and thus this assembly is paraphyletic. Surprisingly, we find substantial phylogenetic conflicts within the plastid genomes of the Poales, as well as among the plastid, mitochondrial, and nuclear data. These conflicts suggest that the Poales could have a complicated evolutionary history, such as rapid radiation and polyploidy, particularly allopolyploidy through hybridization. In sum, our study presents a new perspicacity into the complex phylogenetic relationships and the underlying phylogenetic conflicts within the Poales.

Keywords: Poales, phylogenomic conflict, plastome, mitochondrial, nuclear

OPEN ACCESS

Edited by:
Stefan Wanke,
Technical University Dresden, Germany

Reviewed by:
Ritesh Choudhary,
Agharkar Research Institute, India
Robin Van Velzen,
Wageningen University and Research, Netherlands

***Correspondence:**
De-Zhu Li
dzl@mail.kib.ac.cn
Peng-Fei Ma
mapengfei@mail.kib.ac.cn

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *29 November 2021*
Accepted: *22 December 2021*
Published: *31 January 2022*

Citation:
Wu H, Yang J-B, Liu J-X, Li D-Z and Ma P-F (2022) Organelle Phylogenomics and Extensive Conflicting Phylogenetic Signals in the Monocot Order Poales. Front. Plant Sci. 12:824672. doi: 10.3389/fpls.2021.824672

INTRODUCTION

The order, Poales is a large group of flowering plants in the monocotyledons and belongs to the Commelinid clade, which includes the other three orders of Arecales, Commelinales, and Zingiberales (Angiosperm Phylogeny Group IV [APG IV] et al., 2016). With more than 20,000 species, Poales accounts for about 7% of the angiosperm and 33% of the monocot diversity, respectively (Givnish et al., 2010; Bouchenak-Khelladi et al., 2014; Alves et al., 2015). The species diversity of Poales is extremely uneven among these families. The largest family is Poaceae having about 12,000 species and the smallest one is Ecdeiocoleaceae with only three species

(Christenhusz and Byng, 2016; Hochbach et al., 2018). These species are widely distributed around the world, from the equator to the pole, from floating aquatic plants to the most water-deficient deserts, and most soil types (Stevens, 2001 onward). Moreover, they are generally becoming the dominant species in their ecological communities, such as the grasses (Poaceae) in the savanna and grassland and sedges (Cyperaceae) in the wetland (Linder and Rudall, 2005). Many species of Poales also have significant economic values with Poaceae as the most economically important family in the plant kingdom (Vallée et al., 2016). This family includes many food crops, e.g., rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), and maize (*Zea mays* L.), as well as a variety of bamboos that have multiple applications (Saarela et al., 2018). The pineapple [*Ananas comosus* (L.) Merr.] in Bromeliaceae is a famous tropical fruit and an ornamental plant (Chen et al., 2019). *Typha orientalis* C. Presl and *T. angustifolia* L. from Typhaceae are widely used in weaving and paper industries (Sun, 1992).

Based on a phylogeny using the plastid *rbc*L gene, Duvall et al. (1993) first proposed a composition of 16 families of Poales. Since then, the delineation of Poales has gradually been transformed based on the combined morphological and plastid DNA evidence (Kellogg and Linder, 1995; Angiosperm Phylogeny Group [APG], 1998; Chase et al., 2000; Bremer, 2002). In APG IV, two families, Anarthriaceae and Centrolepidaceae were merged into Restionaceae (Angiosperm Phylogeny Group IV [APG IV] et al., 2016). However, the phylogenetic relationships among them are disputed and the delimitation of the Restionaceae is still problematic (Linder and Rudall, 1993; Bremer, 2002; Michelangeli et al., 2003; Briggs et al., 2014; Hochbach et al., 2018).

The 16 families of Poales can be generally divided into five clades or grades (Linder and Rudall, 2005). The Bromeliaceae, Rapateaceae, and Typhaceae comprise the early diverging grade. The remaining four clades are called "core Poales," which include the cyperid, xyrid, restiid, and graminid clades. The cyperid clade is strongly supported, including Cyperaceae, Juncaceae, and Thurniaceae. As the probable sister group of the cyperid clade, the xyrid clade consists of Eriocaulaceae, Mayacaceae, and Xyridaceae. However, the phylogenetic position and the relationship of the xyrid clade are still ambiguous (Givnish et al., 2018; Hochbach et al., 2018). The Restionaceae, Centrolepidaceae, and Anarthriaceae form the restiid clade and sister to the graminid clade, which encompasses the remaining four families, Ecdeiocoleaceae, Flagellariaceae, Joinvilleaceae, and Poaceae (Hochbach et al., 2018).

The phylogeny of Commelinid has always been a hot topic in the tree of life of monocots and the position of Poales in it has been determined (Barrett et al., 2016). However, a few studies focused on the Poales despite their high ecological and economical significance. The first study focusing on the Poales with a large-scale dataset was provided by Givnish et al. (2010) who sequenced 81 plastid genes of 34 representative species from 15 families. Although the backbone phylogeny of Poales has been reconstructed in this study, there are still many uncertainties about its phylogenetic relationships. First, in contrast to the earliest divergence of Bromeliaceae suggested by Chase et al. (2006),Givnish et al. (2010), Barrett et al. (2016), Givnish et al. (2018), and Hochbach et al. (2018) used the *mat*K to reveal the Bromeliaceae and Typhaceae in an early diverging polytomy and this was supported by a large number of plastid-, mitochondrial- and nuclear-based studies in which Bromeliaceae + Typhaceae were resolved as the early diverging group (Christin et al., 2008; Soltis et al., 2011; Bouchenak-Khelladi et al., 2014; Hertweck et al., 2015; Baker et al., 2021). Moreover, McKain et al. (2016) used the transcriptome data to show that Typhaceae was the first lineage to diverge within the Poales followed by Bromeliaceae. Interestingly, Darshetkar et al. (2019) analyzed the 81 plastid genes dataset like Givnish et al. (2010) with different software and models and obtained the conflicting result with Typhaceae being the first diverging lineage followed by Bromeliaceae. Second, many studies supported Ecdeiocoleaceae as sister to Poaceae (Givnish et al., 2010, 2018; Barrett et al., 2016; Darshetkar et al., 2019; Li et al., 2019). However, the Ecdeiocoleaceae + Joinvilleaceae clade was revealed to be a sister to Poaceae with increased taxon sampling (Bouchenak-Khelladi et al., 2014; McKain et al., 2016; Hochbach et al., 2018; Baker et al., 2021). Third, the relationship involving the Mayacaceae and the xyrid clade is enigmatic. Earlier studies suggested that the Mayacaceae was either within or closely related to the xyrid clade (Michelangeli et al., 2003; Linder and Rudall, 2005; Givnish et al., 2010, 2018; Darshetkar et al., 2019), but other studies suggested that it was within the cyperid clade (Chase et al., 2000, 2006; Janssen and Bremer, 2004) or between the xyrid clade and the cyperid clade (Davis et al., 2004; Hertweck et al., 2015; McKain et al., 2016). Recently, Baker et al. (2021) used 353 nuclear genes to construct the tree of life of angiosperms and found that the Mayacaceae was resolved as an early diverging lineage of Poales, only after the divergence of Bromeliaceae + Typhaceae. In short, these conflicting phylogenetic relationships may be due to the different molecular markers and/or the sparse taxon sampling, as well as various phylogenetic reconstruction methods used in previous studies.

Phylogenomics is an effective way to reconstruct the tree of life based on the genome-scale data (Delsuc et al., 2005; Yu and Zhang, 2006; Zou and Ge, 2008; Zeng et al., 2014; Wen et al., 2015). The data sources of phylogenomics can be either from organelle genomes or from nuclear genomes. In plants, organelle genomes include the plastid genome and mitochondrial genome, which mostly follow a uniparental inheritance (Zeng et al., 2014). In general, the nuclear genome has a faster evolutionary rate and the mitochondrial genome has a slower evolutionary rate while the plastid genome has a moderate rate (Tian and Li, 2002; Wei et al., 2005). The plastid genome is widely used for plant phylogenomic studies (Wei et al., 2005; Leseberg and Duvall, 2009; Gao et al., 2010; Davis et al., 2014; Zeng et al., 2014; Zhao et al., 2021). Plastid phylogenomics has successfully solved a number of plant phylogenetic problems involving taxonomic categories from high to low, e.g., the tree of life of angiosperms at the ordinal level (Li et al., 2019), phylogenetic relationships at the familial level of Malpighiales (Xi et al., 2012), or within the families of Rosaceae and Leguminosae (Zhang et al., 2017, Zhang R. et al., 2020), and the establishment of the phylogenetic framework of the bamboo

tribe, Arundinarieae (Ma et al., 2014, 2017). Compared with the plastid genome, the plant mitochondrial genome has a slower evolutionary rate but a higher genomic rearrangement rate and thus poses challenges for assembly (Wei et al., 2005). These features restrict its application in the study of plant phylogeny. However, some studies have shown that mitochondrial genes could provide additional evolutionary information and are useful in reconstructing the plant phylogeny (Norman and Gray, 2001; Perrotta et al., 2002; Guo and Ge, 2004; Qiu et al., 2010; Bock et al., 2014; Sun et al., 2015; Folk et al., 2016).

In phylogenomics, hundreds to thousands of DNA loci are used, and the phylogenetic discordance or conflicting gene trees appear frequently. The reasons could be the stochastic error and the systematical error, and more often, biological factors including horizontal gene transfer, hybridization, introgression, gene duplication and loss, incomplete lineage sorting, and non-allelic gene conversion (Zou and Ge, 2008; Degnan and Rosenberg, 2009; Sun et al., 2015; Harpak et al., 2017; Kapli et al., 2020). A lot of studies have revealed inconsistencies among plastid, mitochondria, and nuclear phylogenies in plants (Wendel et al., 1995; Sun et al., 2015; Vargas et al., 2017; Hochbach et al., 2018; Jost et al., 2021). In addition, several recent studies suggested a phylogenetic discordance within the plastome at varied evolutionary scales (Gonçalves et al., 2019; Walker et al., 2019; Zhang R. et al., 2020; Yang et al., 2021), questioning the traditional concept of plastomes as a single inherited unit, or at least treating them uncritically in phylogenetic analyses (Gonçalves et al., 2019). More empirical studies are demanded to describe and understand the source, scope, and consequences of conflicting phylogenetic signals in the plastid genome (Yang et al., 2021).

Here, we adopted the classification of 16 families of Poales in this study as the basis for analyses. We reconstructed the phylogeny of Poales with plastid and mitochondrial genomes with at least two taxa or samples for each of the 16 families. The aims of this study are to (1) resolve phylogenetic relationships of key nodes for Poales involving the early diverging grade, xyrid clade, and the sister group of Poaceae; (2) explore the potential conflict within the plastid genome in the Poales phylogeny; and (3) compare the phylogenies built from different plant genomes. With the broadest taxon sampling and the genomic-scale level data, our study resolved ambiguous phylogenetic relationships within the Poales and provided new insights into the conflicting signals among different genomes.

MATERIALS AND METHODS

Plant Materials, Sequencing, Assembly, and Annotation

We selected and sampled representative species in considering the total species of each family of the Poales. All 16 families had a sampling of at least two species. We increased samplings for large families accordingly and 19 species for both Poaceae and Cyperaceae, eight species for Bromeliaceae, four to six species for five families with a total number of 50–1,000 species, and two to four species for eight families less than 50 species. Our sampling could represent the species diversity and phylogenetic diversity of families as far as possible. For the two main systematic problems concerned, we increased the sampling of Bromeliaceae, Typhaceae, Ecodeiocoleaceae, and Joinvilleaceae by two to three species (or individuals).

Illumina sequencing of genomic DNA was undertaken with about 2–10 GB of raw data with 150 bp paired-end reads generated for each sample. Plastomes and mitochondria gene sequences were assembled *de novo* using the GetOrangelle pipeline (Jin et al., 2020). For mitochondria gene sequence, we selected a reference mitochondria genome (*Oryza sativa* L. Indica Group, NC_007886) to blast and retrieve the output contigs by GetOrangelle, many of which were derived from the mitochondrial genome, and further assembled them in Geneious v9.1.4 (Kearse et al., 2012). We also downloaded 33 published plastomes of Poales for analyses, totaling 99 Poales accessions representing 91 species from 50 genera and 16 families. Meanwhile, we selected seven species of Commelinales and Zingiberales as outgroup taxa based on previous studies (Barrett et al., 2016). The corresponding species voucher and GenBank accession numbers are listed in **Supplementary Table 1**.

For the mitochondrial dataset, we obtained sequences of 48 taxa (50 accessions) and downloaded 15 additional complete mitochondrial genomes. As the combination of multi-locus data of representative taxa with single loci from multiple species can generate reliable higher-level phylogenies (Talavera et al., 2021), the *cob* gene sequences of seven species were also included, resulting in a total of 66 species (72 accessions) representing 16 families and 43 genera of Poales. Since there are no published mitochondrial genomes available for the Commelinales and Zingiberales, we selected two species of Arecales [*Phoenix dactylifera* L. (NC_016740) and *Cocos nucifera* L. (NC_031696)] and one species of Asparagales [(*Allium cepa* L. (NC_030100)] as the outgroup. The corresponding species voucher and GenBank accession numbers are listed in **Supplementary Table 2**.

The assembled plastomes were annotated with the PGA software (Qu et al., 2019), followed by manual examination and adjustment in Geneious v9.1.4 (Kearse et al., 2012). Mitochondrial genes were annotated in Geneious v9.1.4 (Kearse et al., 2012) according to the gene annotation information of three grass mitochondrial genomes [*Oryza sativa* Indica Group, NC_007886, *Sorghum bicolor* (L.) Moench, NC_008360 and *Zea mays* L. NB, NC_007982], and mitochondrial genes with more than 80% similarity with reference sequence genome were selected for annotation and tree construction.

Phylogenomic Analysis

We extracted the coding regions in PhyloSuite (Zhang D. et al., 2020), including 80 protein-coding, 4 rRNA and 30 tRNA genes of plastomes, and 28 mitochondrial genes, respectively. We obtained two plastid matrices, 114 genes (114PG), 80 protein-coding genes (80PG), and one mitochondrial matrix of 28 genes (28MG). The nucleotides were first translated into amino acid sequences and aligned with MAFFT v.5 (Katoh et al., 2005) software, and then we used PAL2NAL (Suyama et al., 2006) to obtain the corresponding nucleotide alignment. The ambiguously aligned regions were deleted by Gblocks (Castresana, 2000), and the parameters of allowed gap positions included all, with half, and none for the above three matrices.

TABLE 1 | Characteristics of plastid and mitochondrial matrices used for phylogenetic analyses of Poales.

Matrix	No. of species	No. of genes	Length (bp)	No. of parsimony-informative sites	No. of variable sites	Missing data
114PG	106	114	83,266	31,490	8,233	12.70%
114PG-all	106	114	74,643	30,499	7,231	9.90%
114PG-half	106	114	71,445	29,436	6,832	9.30%
114PG-no	106	105	34,176	13,395	2,887	12.70%
114PG-12	106	114	58,068	22,695	5,878	0.00%
80PG	106	80	75,593	30,287	7,495	13.80%
80PG-all	106	80	67,728	29,349	6,548	10.80%
80PG-half	106	80	64,968	28,407	6,220	10.20%
80PG-no	106	72	31,115	12,908	2,494	1.90%
80PG-12	106	80	50,396	21,492	5,140	13.80%
80PG-3	106	80	25,198	8,795	2,355	13.80%
28MG	75	28	23,709	6,608	2,571	26.40%
28MG-all	75	28	22,383	6,324	2,452	25.40%
28MG-half	75	28	21,704	6,186	2,374	24.90%
28MG-no	75	28	14,529	3,624	1,478	25.30%
80PG_OR_EJ12	106	79	66,179	27,894	6,049	9.70%
80PG-all_OR_EJ12	106	79	60,600	27,135	5,443	7.50%
80PG-half_OR_EJ12	106	79	64,794	28,401	6,219	10.00%
80PG-no_OR_EJ12	106	71	29,775	12,168	2,382	9.70%
80PG_OR_BT123	106	78	74,600	29,986	7,439	14.00%
80PG-all_OR_BT123	106	78	66,735	29,048	6,492	11.00%
80PG-half_OR_BT123	106	78	63,984	28,107	6,164	10.30%
80PG-no_OR_BT123	106	71	30,921	12,857	2,483	2.00%
80PG_OR_EMX123	106	78	58,013	24,020	4,888	2.60%
80PG-all_OR_EMX123	106	78	54,738	23,602	4,647	2.20%
80PG-half_OR_EMX123	106	79	59,706	25,120	5,502	6.10%
80PG-no_OR_EMX123	106	70	27,972	11,895	2,286	2.00%

PG, plastid gene; MG, mitochondrial gene; all, using Gblocks that allow gap positions with all; half, using Gblocks that allow gap positions with a half; no, using Gblocks that allow gap positions with none; −12, 1st + 2nd codon positions of the matrix; −3, 3rd codon; OR, outlier removed.

Matrices of the 1st + 2nd codon positions of 114PG, and 1st + 2nd of 80PG matrix, as well as the 3rd codon positions, were, respectively, obtained by SEAVIEW (Gouy et al., 2009). In total, we acquired 11 plastid matrices and four mitochondrial matrices for concatenate and coalescent analyses (**Table 1**).

For concatenate method, Bayesian inference (BI), maximum likelihood (ML), and maximum parsimony (MP) were employed. ML analyses were conducted in IQ-TREE v.1.6.10 (Nguyen et al., 2015) and RAxML v.8.2.12 (Stamatakis, 2015), respectively. IQ-TREE was performed with Ultrafast bootstrap with 1,000 replicates and the best model (**Supplementary Table 3**), and other default parameters. RAxML was implemented with 1,000 replicates using the GTRGAMMA model and other default parameters. MP analyses were conducted by PAUP*4.0b10 (Cummings, 2004). A heuristic search was executed with 1,000 replicates, random addition, and the tree bisection-reconnection (TBR) branch swapping with the MULTrees option. The bootstrap (BS) method with a heuristic search was performed with 1,000 replicates. BI analyses were implemented using the Mrbayes 3.6.2 (Ronquist et al., 2012) plugin in PhyloSuite (Zhang D. et al., 2020) and the best model with Corrected Akaike information criterion (AICc) was selected by ModelFinder (Kalyaanamoorthy et al., 2017). The Markov chain Monte Carlo (MCMC) algorithm was performed with 2,000,000 (plastid matrices) and 6,000,000 (mitochondrial matrix) generations. Every 1,000 generations sampled one tree with the first 25% of generations abandoned as burnt-in. The trees that remained after reaching a stationary state with the average standard deviation of the split frequencies less than 0.01 were recognized as the consensus trees.

For coalescent analysis, we used ASTRAL-III (Zhang et al., 2018) to infer the species tree with the 80 (80PG matrix) and 72 (80PG-no matrix) plastid gene trees estimated from RAxML with 100 replicates. The branches with BS less than 10% in the gene tree were collapsed using the "nw_ed" code of utilities tool (Junier and Zdobnov, 2010). FigTree v. 1.4.4 was used to visualize the phylogenetic trees (Rambaut, 2012). In general, we defined full support as the posterior probability (PP) = 1.00, BS values = 100%, and local posterior probability (LPP) = 1.00; strong support as PP $\geq$ 0.99, BS $\geq$ 85%, LPP $\geq$ 0.9; moderate support as 0.9 $\leq$ PP $<$ 0.99, 70% $\leq$ BS $<$ 85%, and 0.85 $\leq$ LPP $<$ 0.9; and weak support as PP $<$ 0.9, BS $<$ 70%, and LPP $<$ 0.85.

Quantification Branch Support Values

To further quantify branch support values, we used the Quartet Sampling (QS) method with 1,000 replicates (Pease et al., 2018). This method can also distinguish branches with low

information from those with multiple highly supported but mutually exclusive phylogenetic relationships. Three QS scores were used: (1) Quartet concordance (QC) is the frequency of the concordant quartet inferred over both discordant quartets. (2) Quartet differential (QD) indicates that whether one alternative relationship is sampled more often than the other. (3) Quartet informativeness (QI) is the proportion of replicates that were informative (Pease et al., 2018). The quartet sampling outputs were visualized by the R-script QS_visualization.[1]

Quantification of Phylogenetic Signal for Alternative Tree Topologies

We assessed the phylogenetic signals for three sets of conflicting topologies based on previous studies (Shen et al., 2017, 2021; Zhang R. et al., 2020; Yang et al., 2021). Briefly, we computed the site-wise log-likelihood (SLS) using IQ-TREE and the differences in gene-wise log-likelihood scores (ΔGLS) between conflicting topologies using the Perl scripts of Shen et al. (2021). These analyses allowed us to quantify the phylogenetic signal distribution of alternative topologies at the site and gene levels, and to visualize the proportion of genes that support each topology. The three sets examined were as follows: (a) EJ for the sister relationship of Poaceae, EJ1 of [Poaceae, (Ecdeiocoleaceae, Joinvilleaceae)] vs. EJ2 of [Joinvilleaceae, (Ecdeiocoleaceae, Poaceae)]; (b) BT for the relationship of the early diverging group of Poales, BT1 of [(Typhaceae, Bromeliaceae), (Rapateaceae, core Poales)] vs. BT2 of Typhaceae, [Bromeliaceae, (Rapateaceae, core Poales)] vs. BT3 of Bromeliaceae [Typhaceae, (Rapateaceae, core Poales)]; and (c) EMX for the relationship of Mayacaceae, Eriocaulaceae, and Xyridaceae, EMX1 of EMX1 of Mayacaceae [Eriocaulaceae, (Xyridaceae (restiids, graminids))] vs. EMX2 of (Mayacaceae, Eriocaulaceae), [Cyperids, (Xyridaceae (restiids, graminids))] vs. EMX3 of (Mayacaceae, Eriocaulaceae), [(restiids, graminids), Xyridaceae].

In the analysis of supermatrix, just one or two outlier genes could have a significant effect on the phylogenetic topology (Brown and Thomson, 2016; Shen et al., 2017; Walker et al., 2019; Yang et al., 2021). In order to test this potential effect, we re-built phylogenies after removing the outlier genes in these matrices. The outlier genes were defined as those with phylogenetic signals deviating from a Gaussian-like distribution (Zhang R. et al., 2020; Yang et al., 2021). After removing these genes (**Supplementary Table 5**), we obtained 12 matrices (**Table 1**), i.e., 80PG_outlier_removed_EJ12 (*ycf*2 removed), 80PG-all_outlier_removed_EJ12 (*ycf*2), 80PG-half_outlier_removed_EJ12 (*ycf*2), 80PG-no_outlier_removed_EJ12 (*ndh*F), 80PG_outlier_removed_BT123 (*ycf*3 and *pet*D), 80PG-all_outlier_removed_BT123 (*ycf*3 and *pet*D), 80PG-half_outlier_removed_BT123 (*ycf*3 and *pet*D), 80PG-no_outlier_removed_BT123 (*ycf*3), 80PG_outlier_removed_EMX123 (*ycf*1 and *ycf*2), 80PG-all_outlier_removed_EMX123 (*ycf*1 and *ycf*2), 80PG-half_outlier_removed_EMX123 (*ycf*1), and 80PG-no_outlier_removed_EMX123 (*ndh*A and *psa*B), for analyses.

[1]https://github.com/ShuiyinLIU/QS_visualization

Test of Topological Concordance

We further tested the conflict and the concordance of gene trees and species trees of 80PG and 80PG-no matrices using PhyParts (Smith et al., 2015). We used the Phyx v1.01 (Brown, 2019) to re-root the gene tree and species tree. Meanwhile, the BS support value that is higher than 50% was retained while those lower than 50% did not provide conflict or concordance information. We then separately mapped the 80 and 72 gene trees onto the corresponding species trees by PhyPart. The output of PhyPart was visualized by the script, phypartspiecharts.py (Johnson, 2020).

RESULTS

Taxon Sampling, Plastome, and Mitochondrial Matrix Characteristics

We newly assembled plastomes for 59 species and obtained 53 complete plastomes and 11 plastomes with gaps. The majority of these genomes were annotated to have 68–80 protein-coding genes, 4 rRNA genes, and 16–30 tRNA genes. The two main plastid gene matrices included 114 genes (114PG) and 80 genes (80PG), and the aligned sequences were 83,266 bp and 75,593 bp, respectively. The proportion of missing data ranged from 0.00 to 13.8% for the 11 matrices.

Three complete mitochondrial genomes and 46 at the scaffold level were extracted for 11–28 protein-coding genes for phylogenetic analysis. The alignment of the 28MG matrix was 23,709 bp in length. The proportion of missing data of four matrices ranged from 24.90 to 26.40%. The detailed information of all 15 matrices can be found in **Table 1**.

Plastid Phylogenetic Tree

For ML analyses, phylogenetic trees of Poales constructed using 11 plastid matrices shared a largely consistent topology, except that the relationships among the three families of Mayacaceae, Eriocaulaceae, and Xyridaceae were different (**Figure 1** and **Supplementary Figures 1–10**). The support values of different matrices varied narrowly with generally high support (**Figure 1** and **Supplementary Figures 1–10**). Therefore, the ML topology of the 80 PG-no matrix was selected for discussion, and the BS of RAxML (MLBS) and IQ-TREE (UFBS) were provided on branches.

In the phylogenetic trees of nine plastid matrices, the grouping of Typhaceae and Bromeliaceae was the first lineage diverging within the Poales (**Figure 1** and **Supplementary Figures 1–10**). This node received weak to strong support (BS: 58–90%), and then sister to Rapateaceae and the grouping of the other families with full support. For the remaining two matrices of the 1st + 2nd codon positions of 114PG-12 and 80PG-12, the Bromeliaceae was the earliest diverging lineage with weak support (BS 48–70%).

The cyperid clade received full support in all 11 matrices. In addition, the phylogenetic relationships within it were identical in all the matrices with the Cyperaceae sister to Juncaceae and then the sister to Thurniaceae (**Figure 1** and **Supplementary Figures 1–10**).

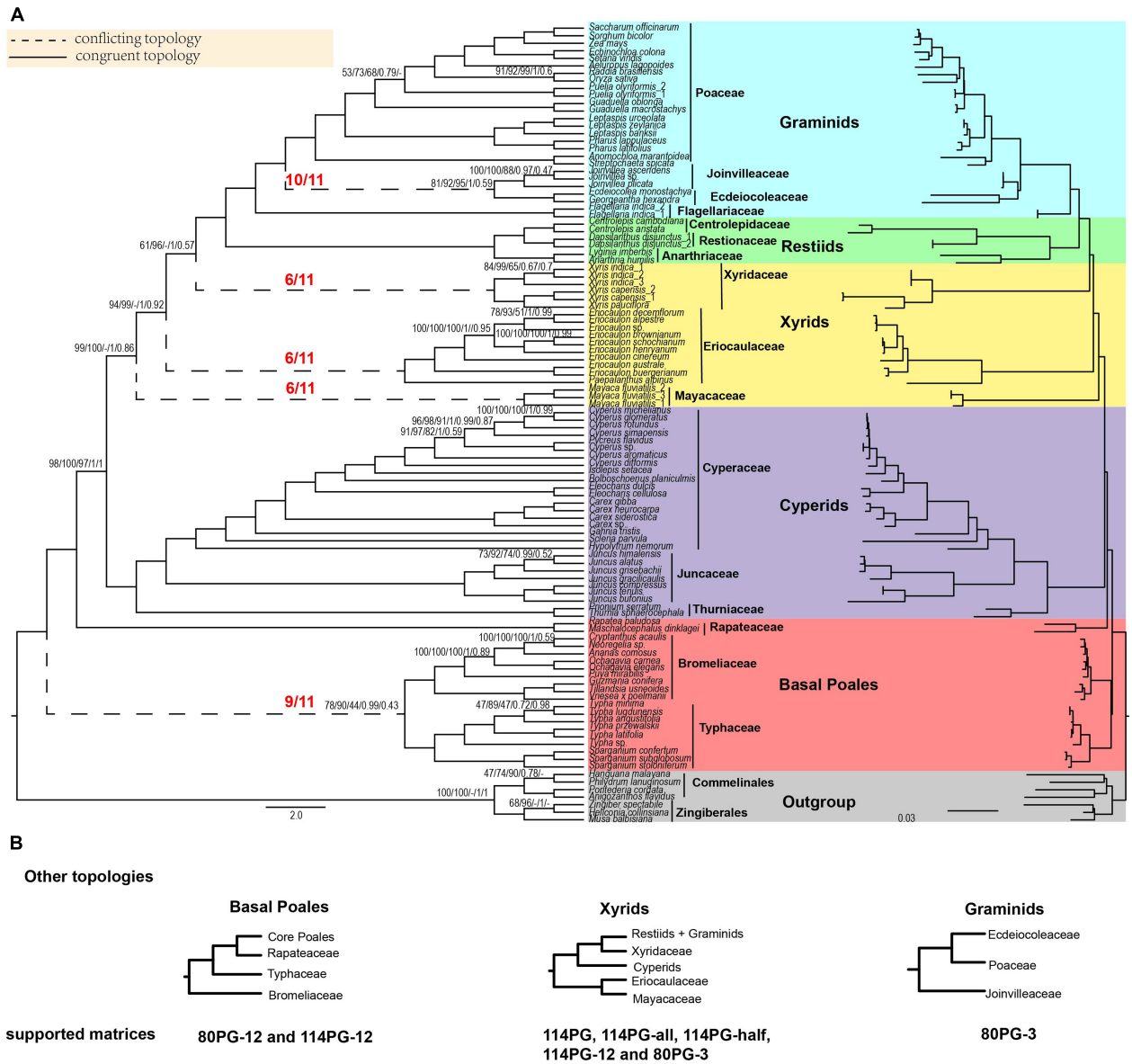


FIGURE 1 | Poales phylogeny based on plastid genes and main conflicting topologies. **(A)** The Poales phylogeny is based on 80 plastid genes using Gblocks with no gap positions allowed (80PG-no). The red fractions on the branches represented by dotted lines indicate the supported numbers in the 11 matrices. Support values are given above branches with the order RAxML bootstrap (MLBS) values/IQ-TREE bootstrap (UFBS) values/MP bootstrap (MPBS)/posterior probability (PP) values/local posterior probability (LPP) values. Support values are only displayed for branches with BS support less than 100%, posterior probability less than 1.00, or local posterior probability less than 1.00. The conflicting topologies are indicated by "–". Different colors represent different clades: Early diverging Poales in red, cyperids in purple, xyrids in yellow, restiids in green, graminids in blue, and the outgroup in gray. **(B)** Other topologies are different from panel **(A)** of basal Poales, xyrids, and graminids recovered in corresponding matrices. Note: -all, -half, and -no mean that using Gblocks with allowing gap positions with all, with half, and none. -12 mean 1st + 2nd condon position of the matrix; -3 mean 3rd codon position of the matrix.

There were two different topologies found for the xyrid clade (**Figure 1** and **Supplementary Figures 1–10**). For the first one, the Eriocaulaceae was the sister to Mayacaceae with a close relationship to the cyperid clade with weak to strong support (BS = 58–99%), and the Xyridaceae became an independent lineage that was close to the restiid clade in five matrices (114PG, 114PG-all, 114PG-half, 114PG-12, and 80PG-3) with full support. For the second one, the xyrid clade was collapsed with Mayacaceae, Eriocaulaceae, and Xyridaceae diverging sequentially along the backbone phylogeny of Poales and the support was BS = 72–99, 61–100, and 100%, respectively.

In the 11 matrices, the restiid clade was all fully supported. The Restionaceae was sister to Centrolepidaceae, and then sister to Anarthriaceae with full support (**Figure 1** and **Supplementary Figures 1–10**). The restiid clade was sister to the graminid clade, which was fully supported in all the 11 matrices. The

Flagellariaceae was sister to the other three families in all matrices with full support. The Ecdeiocoleaceae was sister to Joinvilleaceae in 10 matrices with weak to strong support (BS = 57–92%) and then sister to Poaceae with full support. The 80PG-3 matrix found the grouping of Ecdeiocoleaceae and Poaceae with moderate support (BS = 70–87%) and then sister to Joinvilleaceae with full support (**Figure 1** and **Supplementary Figures 1–10**).

We selected two matrices (80PG and 80PG-no) to run ASTRAL, MP, and BI analyses. For ASTRAL, the early diverging topology of [(Bromeliaceae, Typhaceae) (Rapateacaea, core Poales)] received weak support in 80PG (LPP = 0.61) and 80PG-no (LPP = 0.43). The sister group of Ecdeiocoleaceae and Joinvilleaceae was found in 80PG-no with weak support (LPP = 0.59), while Joinvilleaceae diverged first followed by Ecdeiocoleaceae + Poaceae with moderate support (LPP = 0.86) in 80PG. The other relationships were similar to 80PG-no by ML analyses (**Figure 1** and **Supplementary Figures 11**, **12**).

For MP analyses, the early diverging group of 80PG and 80PG-no was Bromeliaceae + Typhaceae with weak support of maximum parsimony BS (MPBS) = 58 and 44%, respectively, and Ecdeiocoleaceae + Joinvilleaceae was sister to Poaceae with moderate to strong support (MPBS = 74 and 94%). The 80PG matrix found that Eriocaulaceae was sister to Xyridaceae with weak support (MPBS = 53%) and then sister to Mayacaceae with weak support (MPBS = 54%). In the 80PG-no matrix, we obtained a clade of Eriocaulaceae + Mayacaceae with weak support (MPBS = 51%) which was sister to Xyridaceae with strong support (MPBS = 98%). The topology of the other clades was similar to the ML analyses (**Figure 1** and **Supplementary Figures 13**, **14**).

For BI analyses, 80PG and 80PG-no generated the same topology as the ML analyses (114PG-no, 80PG, 80PG-all, 80PG-half, and 80PG-no). The Bromeliaceae + Typhaceae clade was the early diverging group with strong support [posterior probability (PP) = 0.99] in two matrices. The Ecdeiocoleaceae + Joinvilleaceae clade was sister to Poaceae with full support (PP = 1.00) in 80PG-no and weak support (PP = 0.52) in 80PG. The internal relationship of the xyrid clade was Eriocaulaceae, Mayacaceae, and Xyridaceae diverging in sequence. The topology of the other branches was similar to the ML analyses (**Figure 1** and **Supplementary Figures 15**, **16**).

Mitochondrial Phylogenetic Tree

We used four matrices of mitochondrial genes to reconstruct the phylogenetic tree of Poales. In ML analyses, the topologies remained the same except for the phylogenetic relationships involving the three families of Eriocaulaceae, Mayacaceae, and Xyridaceae. In all four matrices, the early diverging group was Typhaceae followed by Bromeliaceae with weak to strong support BS = 72–99% (**Figure 2** and **Supplementary Figures 17**–**22**). The xyrid clade was revealed to be paraphyletic. In the IQ-tree, the Mayacaceae diverged first and was followed by Eriocaulaceae + Xyridaceae in the 28MG and 28MG-all matrices, and the support for these two nodes was weak BS (38–69%) (**Figure 2** and **Supplementary Figure 18**). For the 28MG-half and 28MG-no matrices, the Mayacaceae also diverged first while the relationship between Eriocaulaceae and Xyridaceae was unresolved (**Supplementary Figures 20**, **22**). In RAxML, the Mayacaceae diverged first, and Eriocaulaceae and Xyridaceae successively diverged in only 28MG-no matrix with weak support (MLBS = 26–71%) (**Supplementary Figure 21**) and they were sisters to each other with weak support of MLBS = 34–42% for the other three matrices. The restiid clade also became paraphyletic and the Anarthriaceae (*Anarthria humilis* Nees) was embedded in the Restionaceae with weak to strong support BS = 68–93%. The Restionaceae was sister to the cyperid clade with strong support (BS = 85–97%). Surprisingly, the Centrolepidaceae was also embedded in the Cyperid clade and was sister to Thurniaceae with strong support (BS = 85–97%). The sister relationship between Thurniaceae + Centrolepidaceae and Cyperaceae + Juncaceae received full support. The topology of the graminid clade in all the matrices is consistent with the Flagellariaceae in the basal position and Ecdeiocoleaceae + Joinvilleaceae sister to Poaceae with strong support (BS = 89–100%) except for the 28MG-no matrix with weak support (MLBS = 60%) in RAxML (**Figure 2** and **Supplementary Figures 17**–**22**).

For MP analyses, we only analyzed the 28MG matrix. The early diverging family was the Typhaceae and then Bromeliaceae followed by Rapateaceae with MPBS of 52, 32, and 92% (**Supplementary Figure 23**), respectively. Afterward, the Mayacaceae was sister to Xyridaceae with weak support (MPBS = 68%) and the Eriocaulaceae formed a single clade. The Ecdeiocoleaceae and Joinvilleaceae were sisters with moderate support (MPBS = 84%) and as sisters to Poaceae with strong support (MPBS = 98%). The topologies of other branches were concordant with the RAxML tree of 28MG (**Figure 2** and **Supplementary Figure 23**).

For BI analyses, the topology of 28MG was consistent with the RAxML tree of this matrix (**Figure 2** and **Supplementary Figure 24**). The early diverging families of Typhaceae, Bromeliaceae, and Rapateaceae were separated in turn with PP = 0.83, 0.58, and 1.00. The Ecdeiocoleaceae was sister to Joinvilleaceae and this grouping was then sister to Poaceae and also with full support (**Supplementary Figure 24**).

Quantification of Branch Support Values

We chose two matrices (80PG-no and 28MG) to quantify branch support values. The QC score of $\geq$0.5 was considered to be strong to manifest support among quartets (Pease et al., 2018; Larson et al., 2020). The full support (QC = 1) was obtained for the monophyly of each family (**Supplementary Figure 25**).

In the 80PG-no matrix, we found no support (QC = -0.048) for Bromeliaceae + Typhaceae (**Supplementary Figure 25**). The monophyly of cyperid and restiid was in full support (QC = 1). The sister relationship between Thurniaceae and Cyperaceae + Juncaceae was also in full support (QC = 1). The Mayacaceae, Eriocaulaceae, and Xyridaceae diverged in sequence with no support (QC = -0.34 and -0.2) and weak support (QC = 0.12) for the node placing Xyridaceae sister to restiid + graminid. Within the graminid clade, the Ecdeiocoleaceae + Joinvilleaceae received moderate support (QC = 0.38) while receiving full support (QC = 1) for this grouping as sister to Poaceae. In contrast, the QD = 0 denoted

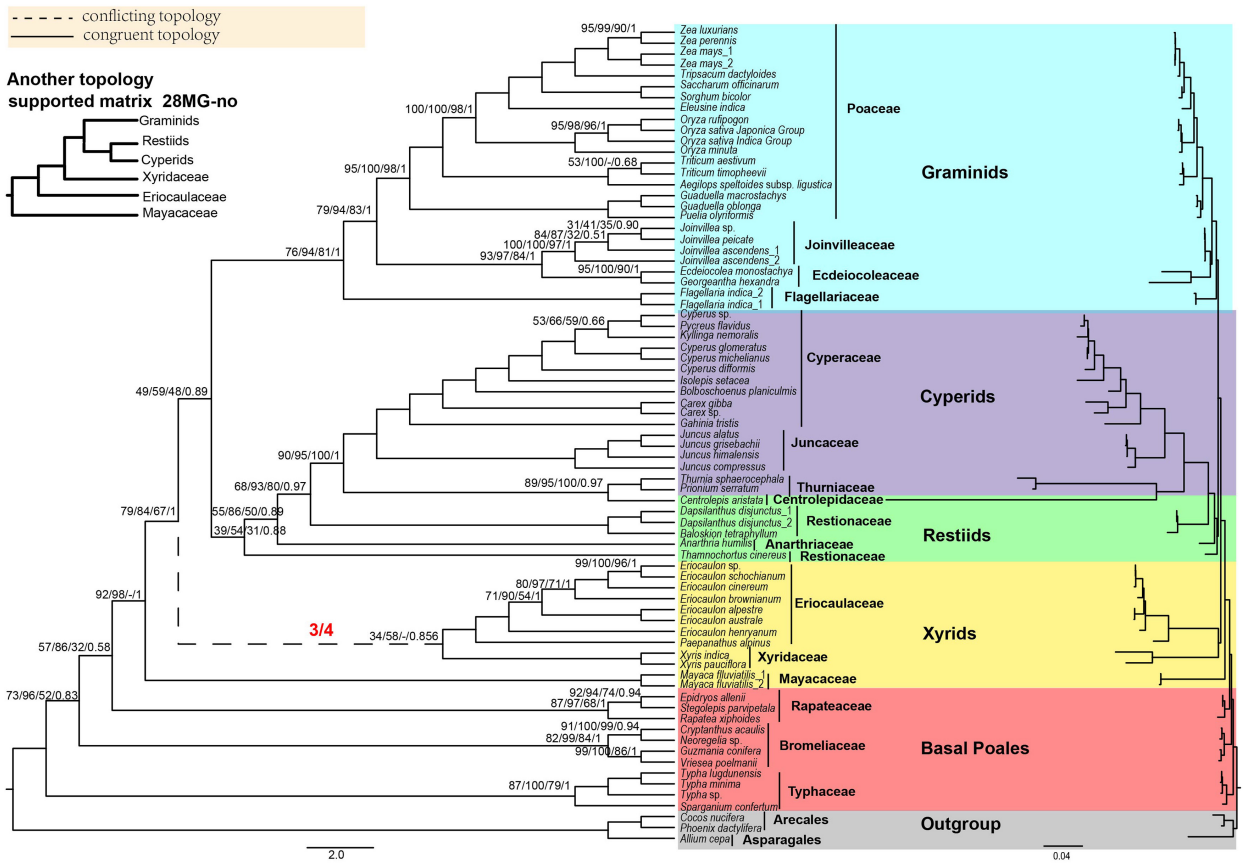


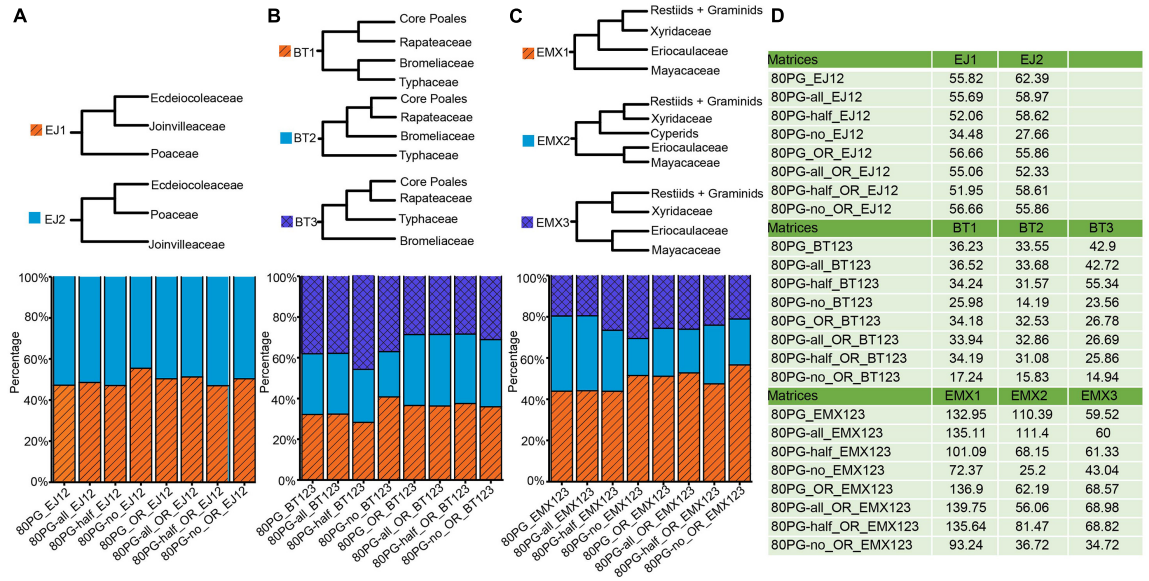
FIGURE 2 | The Poales phylogeny based on 28 mitochondrial genes (28MG). The red fractions on the branches represented by dotted lines indicate the supported numbers in the four matrices. Support values are given above branches with the order MLBS values/UFBS values/MPBS/PP values/LPP values. Support values are displayed for branches with less than 100% BS support, less than 1.00 for posterior probability, or less than 1.00 for local posterior probability. The conflicting topologies are shown by "–". The clades are shown in different colors in **Figure 1**. The top left corner shows another topology that is different from the 28 MG matrix of xyrids recovered in the corresponding matrix. Note: -no mean that using Gblocks that allow gap positions with none.

strong alternative relationships about the nodes within the graminid clade, and in general, the scores of less than 0.3 could be considered as that discordant quartets tend to be heavily skewed toward the conflicting topology (Pease et al., 2018; Larson et al., 2020). The sister groups of Ecdeiocoleaceae and Joinvilleaceae had a low QD score of 0.017 indicating the skew in discordance meaning the possible presence of a supported secondary evolutionary history. Similarly, the nodes connecting Mayacaceae, Eriocaulaceae, and Xyridaceae also received low QD scores (0.27 and 0.044) and thus strong support for alternative evolutionary history. The QD score for the Bromeliaceae + Typhaceae was 0.61, which indicated that inconsistent topologies occurred with relatively equal frequency. The QI scores for all the nodes and the relationships between families were all above 0.76, meaning enough phylogenetic information for these nodes.

For the mitochondrial matrix 28MG, we also obtained no-support (QC = −0.039) for the node placing the Bromeliaceae as the first diverging lineage within the Poales, as well as for the Eriocaulaceae + Xyridaceae (QC = −0.63) (**Supplementary Figure 26**). Moreover, the support was weak for the phylogenetic relationships within the graminid clade with QC scores from 0.13 to 0.37. The QD scores for the nodes connecting Bromeliaceae and Typhaceae and Eriocaulaceae + Xyridaceae were the same as 0.16, indicating a majority of quartets supporting one of the alternative discordant quartet arrangements. The sister relationship between Cyperaceae + Juncaceae and Thurniaceae + Centrolepidaceae had QD = 0 and all the discordant trees sampled were only one of the two alternative topologies. The QI score for the node leading to Typhaceae and Rapateaceae was 0.4, while the other interfamilial relationships were all supported by QI > 0.68.

Quantification of Phylogenetic Signals for Alternative Topologies

We examined phylogenetic signals for three major conflicting topologies for the Poales phylogeny. Phylogenetic signals for the alternative resolutions of each conflicting topology are shown in **Figure 3** and **Supplementary Table 4**. For the conflict involving Ecdeiocoleaceae and Joinvilleaceae, we examined ΔGLS values between EJ1 and EJ2. The proportions of phylogenetical signals



D

Matrices	EJ1	EJ2	
80PG_EJ12	55.82	62.39	
80PG-all_EJ12	55.69	58.97	
80PG-half_EJ12	52.06	58.62	
80PG-no_EJ12	34.48	27.66	
80PG_OR_EJ12	56.66	55.86	
80PG-all_OR_EJ12	55.06	52.33	
80PG-half_OR_EJ12	51.95	58.61	
80PG-no_OR_EJ12	56.66	55.86	
Matrices	**BT1**	**BT2**	**BT3**
80PG_BT123	36.23	33.55	42.9
80PG-all_BT123	36.52	33.68	42.72
80PG-half_BT123	34.24	31.57	55.34
80PG-no_BT123	25.98	14.19	23.56
80PG_OR_BT123	34.18	32.53	26.78
80PG-all_OR_BT123	33.94	32.86	26.69
80PG-half_OR_BT123	34.19	31.08	25.86
80PG-no_OR_BT123	17.24	15.83	14.94
Matrices	**EMX1**	**EMX2**	**EMX3**
80PG_EMX123	132.95	110.39	59.52
80PG-all_EMX123	135.11	111.4	60
80PG-half_EMX123	101.09	68.15	61.33
80PG-no_EMX123	72.37	25.2	43.04
80PG_OR_EMX123	136.9	62.19	68.57
80PG-all_OR_EMX123	139.75	56.06	68.98
80PG-half_OR_EMX123	135.64	81.47	68.82
80PG-no_OR_EMX123	93.24	36.72	34.72

FIGURE 3 | Proportions of phylogenetic signals (ΔGLS) supporting alternative topologies of three conflicting nodes for each of 8 data matrices. **(A)** Proportions of ΔGLS supporting either of two alternative relationships among Ecdeiocoleaceae, Joinvilleaceae, and Poaceae family across eight matrices; **(B)** Proportions of ΔGLS supporting either of the three alternative relationships among Bromeliaceae, Rapateaceae, Typhaceae, and the rest across eight matrices; **(C)** Proportions of ΔGLS supporting either of the three alternative relationships among Eriocaulaceae, Mayacaceae, Xyridaceae, cyperids and the rest across eight matrices; **(D)** the summed ΔGLS values for each matrix. Note: -all, -half, and -no mean that using Gblocks allow gap positions with all, with half, and none. The value −12 means 1st + 2nd condon positions of the matrix; The value −3 means the 3rd codon position of the matrix. The notations _OR and _outlier_removed mean removing outlier genes.

for EJ1 and EJ2 were basically the same (EJ1: 47.04–55.49% vs EJ2: 44.51–52.96%). We found nearly identical proportions of signals of 46.99–51.27% and 49.64–53.01% in the four new matrices for EJ1 and EJ2 after removing the outlier genes (**Supplementary Table 5**), respectively.

For the Bromeliaceae and Typhaceae, computation of ΔGLS values showed ambiguous proportions of sites supporting anyone of the three different topologies of BT1, BT2, and BT3 (**Figure 3**). The proportions of phylogenetical signal for BT1, BT2, and BT3 ranged from 28.26 to 40.77%, from 22.27 to 29.83%, and from 36.97 to 45.68%, respectively. The 80PG-half showed a higher proportion of phylogenetic signals supporting BT3 (45.68%). After removing the outlier genes, we found that the proportions of signals in the four new matrices of BT1, BT2, and BT3 were still low from 35.91 to 37.52%, from 32.97 to 35.15%, and from 28.38 to 31.12%, respectively.

For the relationships among the three families of Eriocaulaceae, Mayacaceae, and Xyridaceae, computation of ΔGLS values also showed ambiguous proportions of sites (**Figure 3**). The higher proportions of phylogenetic signal from 43.84 to 51.47% were shown for EMX1 while lower values from 17.92 to 36.45% and from 19.58 to 30.61% were obtained for EMX2 and EMX3, respectively. We observed increased proportions of signals for EMX1 while decreased values for EMX 2 and EMX3 after removing the outlier genes.

Test of Topological Concordance

The topology of ASTRAL trees of the 80PG and 80PG-no matrix was a bit different (**Supplementary Figure 27**). The one was about the relationship among Eriocaulaceae, Mayacaceae, and Xyridaceae, and another involved Ecdeiocoleaceae and Joinvilleaceae. In 80PG, the Mayacaceae diverged first and Eriocaulaceae and Xyridaceae formed a sister group, while the three families diverged sequentially in 80PG-no. The other was about whether the Joinvilleaceae diverged first (80PG) or grouped with the Ecdeiocoleaceae (80PG-no).

Within the 80 genes of the 80PG matrix, only two genes supported the Bromeliaceae + Typhaceae, but 10 genes against this topology and the remaining genes were uninformative on these relationships. In comparison, none supported and seven genes rejected this topology out of the 72 genes in the 80PG-no matrix. The cyperid clade had 41 genes supported and 15 genes rejected from the 80 genes and a high number of 51 genes supported from the 72 genes, indicating that this clade is stable.

The two nodes connecting Eriocaulaceae, Mayacaceae, and Xyridaceae conflicted with a few supported genes of three and four, respectively, and both had 17 rejected genes in 80PG. In 80PG-no, the Eriocaulaceae + Xyridaceae also only had seven supported genes but 23 rejected genes. In contrast, the restiid clade was stable with high support of 38 and 52 genes in 80PG and 80PG-no, respectively. There were more rejected genes than supported genes for the Ecdeiocoleaceae + Joinvilleaceae (20 vs. 12) in 80PG-no and the first divergence of the Joinvilleaceae (26 vs. 14) in 80PG.

Comparison of Organelle and Nuclear Phylogenies

We further compared our plastid and mitochondrial phylogenies with the recently published nuclear tree of Poales

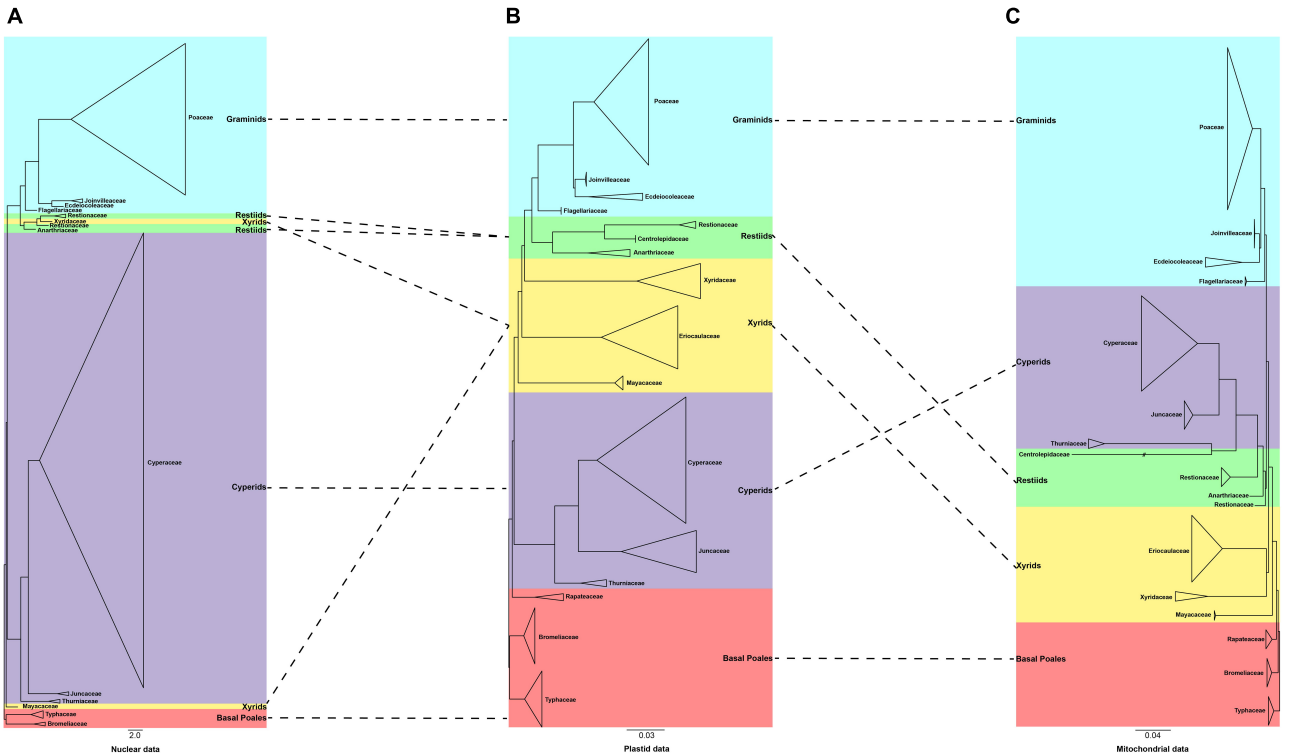


FIGURE 4 | A comparison of the phylogenies obtained from the plastid, mitochondrial, and nuclear matrices. **(A)** 353 nuclear data set [adopted from Baker et al. (2021)] **(B)** 80 plastid data set **(C)** 28 mitochondrial data set. The clades are shown in different colors in **Figure 1**. The branch with "//" was shortened by 50%.

(Baker et al., 2021; **Figure 4**). We found that the position of the graminid remained unchanged. The conflict concentrated on the early diverging group, the cyperid and restiid clades, and the xyrid assembly. Taking the plastid tree as a reference, we explored the conflict in detail. In the plastid phylogeny, the phylogenetic placement of the families of the xyrid assembly is between the cyperid clade and the restiid clade, while in the mitochondrial phylogeny, the position changes and is located between the early diverging taxa and the cyperid clade. For the nuclear data, the placement of the Mayacaceae is the same as that of mitochondria and the Xyridaceae is clustered into the Restionaceae (Baker et al., 2021). The cyperid clade has a close relationship with the early diverging group in the plastid data, while it is located between the xyrid assembly, Mayacaceae and restiid clade in the nuclear data (Baker et al., 2021), and between the graminid clade and the restiid clade in the mitochondrial data. The restiid clade is sister to the graminid clade in the plastid and nuclear data while sister to the cyperid clade in the mitochondrial data.

DISCUSSION

The Early Diverging Poales

The early diverging Poales include three families: Bromeliaceae, Rapateaceae, and Typhaceae. However, the relationships among them are variable in recent studies. The first divergence of Bromeliaceae was supported by the analyses of 75, 77, or 81 plastid genes (Givnish et al., 2010, 2018; Barrett et al., 2016) and the combination of one mitochondrial gene, two rDNA genes, and four plastid genes (Chase et al., 2006). In contrast, Typhaceae was estimated as the first diverging lineage followed by Bromeliaceae based on the analyses of 81 plastid trees in other studies (Darshetkar et al., 2019; Li et al., 2019) or nuclear genes (McKain et al., 2016). Moreover, the sister relationship between Bromeliaceae and Typhaceae was also revealed in some studies, such as using the *rbc*L and *ndh*F genes (Christin et al., 2008; Bouchenak-Khelladi et al., 2014), the low-copy nuclear gene *PHYC* (Hertweck et al., 2015), and 353 nuclear genes (Baker et al., 2021). Analyses of the concatenated plastid genes all uncovered the sister relationship of Bromeliaceae and Typhaceae with weak to strong support (**Figure 1** and **Supplementary Figures 1**–**10**). Similarly, the coalescent ASTRAL analyses also supported Bromeliaceae as sister to Typhaceae despite weak support (**Supplementary Figures 11**, **12**).

Nevertheless, the mitochondrial trees revealed the Typhaceae as the earliest diverging group with weak to strong support (**Figure 2** and **Supplementary Figures 17**–**24**), and this result is in conflict with the plastid phylogeny. The conflict between plastid and mitochondrial phylogenies may be due to the evolutionary histories of these two subcellular compartments being unlinked (Sun et al., 2015) and/or incomplete lineage sorting (Lee et al., 2018). The branches leading to the Bromeliaceae and Typhaceae are very short, indicating that they may endure extreme changes in the rates of molecular evolution

(Givnish et al., 2018). Moreover, the short internal branches and conflicting topology of Typhaceae and Bromeliaceae in trees could also be due to rapid radiation as suggested before (hence lack of sufficient phylogenetic signals) (Barrett et al., 2013; Hertweck et al., 2015; Hochbach et al., 2018). Taken together, we consider the phylogenetic relationship between Bromeliaceae and Typhaceae as polytomy Hochbach et al. (2018) and as the earliest diverging lineages of Poales followed by Rapateaceae.

The Non-monophyly of the Xyrid Assembly

The phylogenetic relationships among Eriocaulaceae, Mayacaceae, and Xyridaceae within the xyrid clade remain unresolved in previous studies. The Eriocaulaceae was placed as sister to Xyridaceae (Christin et al., 2008) or Mayacaceae (Givnish et al., 2010) in analyses using plastid genes, while the recent plastome study recognized the topology of [Mayacaceae, (Eriocaulaceae, (Xyridaceae, (restiids, graminids)))] (Li et al., 2019). Using 353 nuclear genes, Baker et al. (2021) suggested that the Mayacaceae was an early diverging lineage within the Poales following the divergence of Bromeliaceae and Typhaceae, while the Xyridaceae was embedded in the restiid clade.

Five of our concatenated plastid datasets (114PG, 114PG-all, 114PG-half, 114PG-12, and 80PG-3) revealed Eriocaulaceae as sister to Mayacaceae with a close relationship to the cyperid clade, and the Xyridaceae became an independent clade that was close to the restiid clade (**Supplementary Figures 1–3**, **5**, and **10**). The other six matrices supported the topology [Mayacaceae, (Eriocaulaceae, (Xyridaceae, (restiids, graminids)))] (**Figure 1** and **Supplementary Figures 4**, **6–9**). The two conflicting topologies are obtained probably, as a result, the tRNA and rRNA genes have different evolutionary histories from the protein-coding genes in the plastid genome, e.g., faster substitution rate. The third codon positions can mutate more frequently than the first and second positions and thus may experience mutation saturation leading to the phylogenetic artifact (Zeng et al., 2017). In comparison, the ASTRAL analyses revealed the topology [Mayacaceae, ((Eriocaulaceae, Xyridaceae), (restiids, graminids))] and [Mayacaceae, (Eriocaulaceae, (Xyridaceae, (restiids, graminids)))] with weak and strong support in 80PG and 80PG-no, respectively (**Supplementary Figures 11**, **12**). For the mitochondrial data, two different topologies emerged, [Mayacaceae, (Eriocaulaceae, (Xyridaceae, (restiids, graminids)))] and [Mayacaceae, ((Eriocaulaceae, Xyridaceae), (restiids, graminids))]. This result is likely due to the insufficient informative sites in the mitochondrial genes. Whatever be the final phylogenetic resolution of the xyrid clade, it appears that the monophyly of this clade would not be achieved and we can consider it paraphyletic. However, for the convenience of communication, we suggest tentatively keeping the name and calling it the paraphyletic xyrid assembly.

The Restiid Clade

The internal relationship of the restiid clade is still doubtful (Briggs et al., 2014; Hochbach et al., 2018). Previous studies based on the plastid, nuclear, or combined plastid, nuclear, and mitochondrial data (Chase et al., 2006; Bouchenak-Khelladi et al., 2014; Hochbach et al., 2018; Baker et al., 2021) supported a monophyletic Anarthriaceae, which was sister to the Restionaceae. However, the relationship between Centrolepidaceae and Restionaceae had two different topologies. The sister relationship between Centrolepidaceae and Restionaceae was supported by a majority of previous studies with plastid, nuclear, and mitochondrial data (e.g., Chase et al., 2006; Givnish et al., 2010; Bouchenak-Khelladi et al., 2014; McKain et al., 2016; Darshetkar et al., 2019; Li et al., 2019; Baker et al., 2021). In other studies, the Centrolepidaceae was embedded in Restionaceae based on *rbc*L and *atp*B (Bremer, 2002) and the combined data (*mat*K, *Phy*B, and *Topo6*) (Hochbach et al., 2018). In our study, the topology of [Anarthriaceae, (Restionaceae, Centrolepidaceae)] is revealed by plastid data with strong support in both the concatenate and coalescent methods.

However, in our mitochondrial data, the Anarthriaceae is embedded in the Restionaceae. This result may be due to the fact that only one mitochondrial gene of *cob* was available for *Thamnochortus cinereus* H.P. Linder in analyses and this gene is short of phylogenetic information. Moreover, the Centrolepidaceae is placed in the cyperid clade and as a sister to the Thurniaceae with strong support. The delimitation of the Restionaceae and the placement of the Centrolepidaceae has thus not been fully resolved, and more molecular data and taxon samplings are needed for further analysis.

The Graminid Clade

The monophyly of the graminid clade with the Flagellariaceae as the first diverging group is supported by all previous studies (e.g., Chase et al., 2006; Christin et al., 2008; Soltis et al., 2011; Bouchenak-Khelladi et al., 2014; Hertweck et al., 2015; McKain et al., 2016; Hochbach et al., 2018; Darshetkar et al., 2019; Li et al., 2019; Baker et al., 2021). However, the relationship within the remaining families of Ecdeiocoleaceae, Joinvilleaceae, and Poaceae is still blurry. Both the topologies of [Joinvilleaceae, (Ecdeiocoleaceae, Poaceae)] and [(Joinvilleaceae, Ecdeiocoleaceae), Poaceae] were revealed in previous studies (Givnish et al., 2010; Bouchenak-Khelladi et al., 2014; Barrett et al., 2016; McKain et al., 2016; Hochbach et al., 2018; Darshetkar et al., 2019; Li et al., 2019; Baker et al., 2021). Our study supports the latter topology using the concatenate method, as well as the coalescent analysis of the 80PG. In contrast, the coalescent analysis of the 80PG-no generates the topology [Joinvilleaceae, (Ecdeiocoleaceae, Poaceae)] with moderate support. The conflict between coalescent and concatenate methods could be caused by the limitations of ASTRAL when largely uninformative loci exist as in the 80PG-no (Yang et al., 2021). The mitochondrial data also support the topology of [(Joinvilleaceae, Ecdeiocoleaceae), Poaceae] with high support values and we advocate a topology of [Flagellariaceae, ((Joinvilleaceae, Ecdeiocoleaceae), Poaceae)] for the graminid clade.

Conflicting Signals in Plastid Phylogeny

Based on branch support value and phylogenetic signal analyses, we observed the extensive presence of conflicts among plastid loci involving several long-questioned nodes

(e.g., the relationship of Bromeliaceae and Typhaceae, Joinvilleaceae and Ecdeiocoleaceae, and the xyrid assembly). The no-support means a majority of quartets favor one of the alternative discordant quartet arrangement history (Pease et al., 2018). We found no support for the topologies [(Bromeliaceae, Typhaceae), (Rapateaceae, core Poales)] and [Mayacaceae, (Eriocaulaceae, (Xyridaceae, (graminids, restiids)))]. The ΔGLS values show almost the same proportions of sites supporting three kinds of topologies referring to the Bromeliaceae and Typhaceae, illustrating the conflict with three different topologies, and their relationships may be better treated as polytomy at present. In addition, the topology of [Mayacaceae, (Eriocaulaceae, (Xyridaceae, (graminids, restiids)))] and [(Joinvilleaceae, Ecdeiocoleaceae), Poaceae] both have low QD scores, suggesting that there is a skew between two inconsistent topologies. This result points to a given branch having a biased biological process other than the background lineage sorting, including confounding variables, such as introgression, high heterogeneity of evolutionary rate, heterogeneous base composition, etc. (Pease et al., 2018).

The ΔGLS values show the higher proportions of phylogenetic signal for [Mayacaceae, (Eriocaulaceae, (Xyridaceae, (graminids, restiids)))], and the maximum is 47.5%, but the other two topologies have about 50%. Previous studies have suggested that removing problematic sequences would avoid artifacts to some extent and contribute to the robustness of phylogenetic results (Goremykin et al., 2010; Parks et al., 2012; Yang et al., 2021), although identifying these sequences is difficult (Som, 2015). We removed the outlier genes but found that they have little effect on resolving these long-debated relationships between Mayacaceae, Eriocaulaceae, and Xyridaceae. Recently, Doyle (2021) stressed that the plastid genome should be treated as a single unit for phylogenetic analyses. However, some studies indicate that it is not proper to treat the plastid genome as a single unit particularly when the evidence of recombination is present (Gonçalves et al., 2020; Daniell et al., 2021; Yang et al., 2021). Several studies have shown that the conflicts within the plastid genome (Walker et al., 2019; Zhang R. et al., 2020; Yang et al., 2021), and our results also support this point. However, the cause of the plastid conflict has not yet been determined. Stochastic errors, such as those associated with rapid radiation and limited phylogenetic signals may explain the majority of the observed conflicts (Zhang R. et al., 2020; Yang et al., 2021). In plastid phylogenetic analyses, the vast majority of plastid loci are generally uninformative, and a few genes with strong signals will largely determine the phylogenetic resolution as shown here (**Supplementary Figure 27**; Walker et al., 2019; Zhang R. et al., 2020; Yang et al., 2021). The biological cause for the conflict of plastids could be heterogeneous recombination and a gene transfer between genomes (Walker et al., 2019; Zhang R. et al., 2020; Daniell et al., 2021; Yang et al., 2021). However, we did not find any clues for these factors playing a role here. Instead, the phylogenetic tree of the Poales harbors numerous contrasting short and long branches, meaning highly heterogeneous plastome evolutionary rates among these families of the Poales, and this could be one of the reasons for causing conflicts within the plastid genomes (Barrett et al., 2013, Barrett et al., 2016). The species of Poales have a variety of habitats, and ten families of them grow in swamp or wet habitats, among which Typhaceae and Mayacaceae are typical aquatic plants (Linder and Rudall, 2005). The plastome of aquatic plants may have a more complex structure, and the deletion and inversion were found in a large number of aquatic plants, e.g., *Eleocharis* (Cyperaceae) and *Najas flexilis* (Willd.) Rostk. and W.L.E. Schmidt (Hydrocharitaceae) (Peredo et al., 2013; Lee et al., 2020). This may also contribute to the phylogenetic conflict. In addition, the photosynthetic pathways of Poales are also diverse, including all known three pathways of C_3, C_4, and Crassulacean acid metabolism (CAM) (Linder and Rudall, 2005). This is a potential explanation for the high heterogeneity of substitution rates heterogeneity, which ultimately results in phylogenetic conflict (Barrett et al., 2016).

Conflicts Among Three Plant Genomes

The phylogenetic conflict between organelles (plastid and mitochondrial) and nuclear genes has been reported in various taxa, such as coralline red algae, *Lachemilla* (Rosaceae), and *Pterocarya* (Juglandaceae) (Lee et al., 2018; Morales-Briones et al., 2018; Mu et al., 2020). The plastid and mitochondrial genomes have a strong conflict in our study. The mitochondrial data have more missing data and lower coverage than the plastid data, which may be one of the factors causing the conflict between mitochondrial and plastid phylogenies. Another possible explanation is that the mitochondrial genomes of Poales have undergone extensive horizontal gene transfer between nuclear and plastid genomes, which is typical in land plants (Folk et al., 2016). In our study, we found that the systematic position of Centrolepidaceae is extraordinary and the branches are very long (**Figure 2** and **Supplementary Figures 19–23**). This phenomenon is also observed by Folk et al. (2016), who indicates that the foreign DNA of the nucleus may be the cause for the large difference in branch length and location in mitochondrial analysis.

Meanwhile, strong conflicts are also detected between the organelle genomes and nuclear genes. The phylogenetic relationships of the three clades/assembly (cyperid, restiid, and xyrid) are different among the three genomes. This may be caused by insufficient sampling of nuclear data. Three families of Centrolepidaceae, Rapateaceae, and Eriocaulaceae are not sampled in our nuclear data, and the sampling distribution is uneven between each family. For example, the Xyridaceae have only one species included, which may be the reason for the odd phylogenetic position of this family. There is also a potential biological source of the incongruence among plastid, mitochondrial, and nuclear loci (Sun et al., 2015). The factors could be incomplete lineage sorting, hybridization, lateral transfer of organellar genomes, plastome capture, and polyploidy (Stegemann et al., 2012; Lee et al., 2018; Morales-Briones et al., 2018).

In particular, polyploidy is prevalent in plant groups and all the families of Poales have experienced polyploidization events in their evolutionary history (McKain et al., 2016; Van de Peer et al., 2017; Morales-Briones et al., 2018; Wu et al., 2020; Guo et al., 2021). Here, the sampled taxa of *Centrolepis aristata* (R.Br.)

Roem. and Schult. (restiid) and the species of Eriocaulaceae (xyrid) are probably hexaploid (Šmarda et al., 2014), and *Juncus bufonius* L. (cyperid) is octoploid (Kubešová et al., 2010). Meanwhile, hybridization is ubiquitous in the green plant (Soltis and Soltis, 2009; Triplett et al., 2010), and the combination of hybridization and polyploid events (Soltis and Soltis, 2009) is another usual cause of phylogenetic conflict. The conflicts of three genomic data reflected in the branch lengths (Jost et al., 2021) indicate that the molecular evolution rate between different genomes is highly heterogeneous with certain families of the Poales experiencing an accelerated rate of sequence evolution (Guisinger et al., 2010; Barrett et al., 2013, Barrett et al., 2016). The fast-evolving sites are more likely to be saturated and prone to the accumulation of non-phylogenetic signals (Rodríguez-Ezpeleta et al., 2007), and thus leading to topological conflicts among the three genomes. In short, the conflict of three genomic data might be due to a combination of the missing data in mitochondria, polyploid history, heterogeneity of molecular evolution rate, and sparse sampling of nuclear data, and these factors deserve to be explored in detail in future studies. Furthermore, due to the limitations of the organelle genome, the nuclear genomic data will be used to finally resolve the phylogenetic relationships of Poales, so as to improve the understanding of the cause of phylogenetic conflict in this order.

CONCLUSION

With a broad taxon sampling, we used plastid and mitochondria genomes to infer the phylogenetic tree of Poales and found the long-standing controversial nodes of Poales mainly caused by extensive conflict across genomic compartments. For the xyrid assembly, we found it paraphyletic, and its relationship with the three families, Eriocaulaceae, Mayacaceae, and Xyridaceae within the Poales is still not fully resolved. Our study has not only revealed phylogenetic conflicts within the plastid genomes, but also extensive conflicts among the plastid, mitochondrial and nuclear data in the Poales. Many factors, such as the missing data of mitochondrion, insufficient nuclear sampling, rapid radiation, heterogeneity of molecular evolution rate, and allopolyploidy by hybridization are potentially involved in generating these conflicts in the Poales.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

P-FM, D-ZL, and HW conceptualized the study. HW, P-FM, J-BY, and J-XL were involved in data generation and methodology. HW worked on data analysis and visualization. HW, P-FM, D-ZL, J-BY, and J-XL reviewed and revised the first draft of the manuscript and approved the submitted version.

FUNDING

This study was supported by the National Natural Science Foundation of China (Project No. 31770239), CAS's large-scale scientific facilities (Grant No. 2017-LSF-GBOWS-02) and the Program of Science and Technology Talents Training in Yunnan Province (202105AC160022).

ACKNOWLEDGMENTS

We thank the Germplasm Bank of Wild Species in the Southwest China and Molecular Biology Experiment Center, National Wild Plant Germplasm Resource Center for providing convenience in our lab work. We sincerely thank the herbarium of the Royal Botanic Gardens, Kew for herbarium materials. We are grateful to Rong Zhang for the discussion on the quantification of phylogenetic signals. We also thank Xiao-Gang Fu for the ASTRAL-III analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.824672/full#supplementary-material

REFERENCES

Alves, M., Trovó, M., Forzza, R. C., and Viana, P. (2015). Overview of the systematics and diversity of Poales in the Neotropics with emphasis on the Brazilian flora. *Rodriguésia* 66, 305–328. doi: 10.1590/2175-7860201566203

Angiosperm Phylogeny Group [APG] (1998). An ordinal classification for the families of flowering plants. *Ann. Missouri Bot. Gard.* 85, 531–553. doi: 10.2307/2992015

Angiosperm Phylogeny Group IV [APG IV], Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., et al. (2016). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* 181, 1–20. doi: 10.1111/boj.12385

Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., et al. (2021). A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Syst. Biol.* syab035. doi: 10.1093/sysbio/syab035 [Epub ahead of print].

Barrett, C. F., Baker, W. J., Comer, J. R., Conran, J. G., Lahmeyer, S. C., Leebens-Mack, J. H., et al. (2016). Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. *New Phytol.* 209, 855–870. doi: 10.1111/nph.13617

Barrett, C. F., Davis, J. I., Leebens-Mack, J., Conran, J. G., and Stevenson, D. W. (2013). Plastid genomes and deep relationships among the commelinid monocot angiosperms. *Cladistics* 29, 65–87. doi: 10.1111/j.1096-0031.2012.00418.x

Bock, D. G., Kane, N. C., Ebert, D. P., and Rieseberg, L. H. (2014). Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. *New Phytol.* 201, 1021–1030. doi: 10.1111/nph.12560

Bouchenak-Khelladi, Y., Muasya, A. M., and Linder, H. P. (2014). A revised evolutionary history of Poales: origins and diversification. *Bot. J. Linn. Soc.* 175, 4–16. doi: 10.1111/boj.12160

Bremer, K. (2002). Gondwanan evolution of the grass alliance of families (Poales). *Evolution* 56, 1374–1387. doi: 10.1111/j.0014-3820.2002.tb01451.x

Briggs, B. G., Marchant, A. D., and Perkins, A. J. (2014). Phylogeny of the restiid clade (Poales) and implications for the classification of Anarthriaceae, Centrolepidaceae and Australian Restionaceae. *Taxon* 63, 24–46. doi: 10.12705/631.1

Brown, J. M., and Thomson, R. C. (2016). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Syst. Biol.* 66, 517–530. doi: 10.1093/sysbio/syw101

Brown, W. J. (2019). *Phyx v1.01*. Available online at: https://github.com/FePhyFoFum/phyx/releases/tag/v1.01 (accessed September 27, 2019).

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Chase, M. W., Soltis, D. E., Soltis, P. S., Rudall, P. J., Fay, M. F., Hahn, W. H., et al. (2000). "Higher-level systematics of the monocotyledons: an assessment of current knowledge and a new classification," in *Monocots: Systematics and Evolution*, eds K. L. Wilson and D. A. Morrison (Collingwood, ON: CSIRO), 3–16. doi: 10.1201/9781315166339-2

Chase, M., Fay, M., Devey, D. S., Maurin, O., Rønsted, N., Davies, T., et al. (2006). Multigene analyses of monocot relationships. *Aliso* 22, 63–75. doi: 10.5642/ALISO.20062201.06

Chen, L.-Y., VanBuren, R., Paris, M., Zhou, H., Zhang, X., Wai, C. M., et al. (2019). The bracteatus pineapple genome and domestication of clonally propagated crops. *Nat. Genet.* 51, 1549–1558. doi: 10.1038/s41588-019-0506-8

Christenhusz, M. J., and Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa* 261, 201–217.

Christin, P.-A., Salamin, N., Muasya, A. M., Roalson, E. H., Russier, F., and Besnard, G. (2008). Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. *Mol. Biol. Evol.* 25, 2361–2368. doi: 10.1093/molbev/msn178

Cummings, M. P. (2004). "PAUP* "Phylogenetic analysis using parsimony (and other methods)"," in *Dictionary of Bioinformatics and Computational Biology*, eds J. M. Hancock and M. J. Zvelebil (Hoboken, NJ: Wiley).

Daniell, H., Jin, S., Zhu, X. G., Gitzendanner, M. A., Soltis, D. E., and Soltis, P. S. (2021). Green giant—a tiny chloroplast genome with mighty power to produce high-value proteins: history and phylogeny. *Plant Biotechnol. J.* 19, 430–447. doi: 10.1111/pbi.13556

Darshetkar, A. M., Datar, M. N., Tamhankar, S., Li, P., and Choudhary, R. K. (2019). Understanding evolution in Poales: insights from Eriocaulaceae plastome. *PLoS One* 14:e0221423. doi: 10.1371/journal.pone.0221423

Davis, C. C., Xi, Z., and Mathews, S. (2014). Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. *BMC Biol.* 12:11. doi: 10.1186/1741-7007-12-11

Davis, J. I., Stevenson, D. W., Petersen, G., Seberg, O., Campbell, L. M., Freudenstein, J. V., et al. (2004). A phylogeny of the monocots, as inferred from rbcL and atpA sequence variation, and a comparison of methods for calculating jackknife and bootstrap values. *Syst. Bot.* 29, 467–510. doi: 10.1600/0363644041744365

Degnan, J. H., and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340. doi: 10.1016/j.tree.2009.01.009

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375. doi: 10.1038/nrg1603

Doyle, J. J. (2021). Defining coalescent genes: Theory meets practice in organelle phylogenomics. *Syst. Biol.* syab053. doi: 10.1093/sysbio/syab053

Duvall, M. R., Clegg, M. T., Chase, M. W., Clark, W. D., Kress, W. J., Hills, H. G., et al. (1993). Phylogenetic hypotheses for the monocotyledons constructed from rbcL sequence data. *Ann. Missouri Bot. Gard.* 80, 607–619. doi: 10.2307/2399849

Folk, R. A., Mandel, J. R., and Freudenstein, J. V. (2016). Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Syst. Biol.* 66, 320–337. doi: 10.1093/sysbio/syw083

Gao, L., Su, Y.-J., and Wang, T. (2010). Plastid genome sequencing, comparative genomics, and phylogenomics: current status and prospects. *J. Syst. Evol.* 48, 77–93. doi: 10.1111/j.1759-6831.2010.00071.x

Givnish, T. J., Ames, M., McNeal, J. R., McKain, M. R., Steele, P. R., Depamphilis, C. W., et al. (2010). Assembling the tree of the monocotyledons: plastome sequence phylogeny and evolution of Poales1. *Ann. Missouri Bot. Gard.* 97, 584–616. doi: 10.3417/2010023

Givnish, T. J., Zuluaga, A., Spalink, D., Soto Gomez, M., Lam, V. K. Y., Saarela, J. M., et al. (2018). Monocot plastid phylogenomics, timeline, net rates of species diversification, the power of multi-gene analyses, and a functional model for the origin of monocots. *Am. J. Bot.* 105, 1888–1910. doi: 10.1002/ajb2.1178

Gonçalves, D. J. P., Simpson, B. B., Ortiz, E. M., Shimizu, G. H., and Jansen, R. K. (2019). Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. *Mol. Phylogenet. Evol.* 138, 219–232. doi: 10.1016/j.ympev.2019.05.022

Gonçalves, D. J., Jansen, R. K., Ruhlman, T. A., and Mandel, J. R. (2020). Under the rug: abandoning persistent misconceptions that obfuscate organelle evolution. *Mol. Phylogenet. Evol.* 151:106903. doi: 10.1016/j.ympev.2020.106903

Goremykin, V. V., Nikiforova, S. V., and Bininda-Emonds, O. R. P. (2010). Automated removal of noisy data in phylogenomic analyses. *J. Mol. Evol.* 71, 319–331. doi: 10.1007/s00239-010-9398-z

Gouy, M., Guindon, S., and Gascuel, O. (2009). SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27, 221–224. doi: 10.1093/molbev/msp259

Guisinger, M. M., Chumley, T. W., Kuehl, J. V., Boore, J. L., and Jansen, R. K. (2010). Implications of the plastid genome sequence of Typha (Typhaceae, Poales) for understanding genome evolution in *Poaceae*. *J. Mol. Evol.* 70, 149–166. doi: 10.1007/s00239-009-9317-3

Guo, C., Ma, P. F., Yang, G. Q., Ye, X. Y., Guo, Y., Liu, J. X., et al. (2021). Parallel ddRAD and genome skimming analyses reveal a radiative and reticulate evolutionary history of the temperate bamboos. *Syst. Biol.* 70, 756–773. doi: 10.1093/sysbio/syaa076

Guo, Y. L., and Ge, S. (2004). The utility of mitochondrial nad1 intron in phylogenetic study of Oryzeae with reference to the systematic position of Porteresia. *Acta Phytotaxon. Sin.* 42, 333–344.

Harpak, A., Lan, X., Gao, Z., and Pritchard, J. K. (2017). Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proc. Natl. Acad. Sci. U S A.* 114, 12779–12784. doi: 10.1073/pnas.1708151114

Hertweck, K. L., Kinney, M. S., Stuart, S. A., Maurin, O., Mathews, S., Chase, M. W., et al. (2015). Phylogenetics, divergence times and diversification from three genomic partitions in monocots. *Bot. J. Linn. Soc.* 178, 375–393. doi: 10.1111/boj.12260

Hochbach, A., Linder, H. P., and Röser, M. (2018). Nuclear genes, matK and the phylogeny of the Poales. *Taxon* 67, 521–536. doi: 10.12705/673.5

Janssen, T., and Bremer, K. (2004). The age of major monocot groups inferred from 800+rbcL sequences. *Bot. J. Linn. Soc.* 146, 385–398. doi: 10.1111/j.1095-8339.2004.00345.x

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 21:241. doi: 10.1186/s13059-020-02154-5

Johnson, M. (2020). Phypartspiecharts. https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts (accessed November 2, 2021).

Jost, M., Samain, M.-S., Marques, I., Graham, S. W., and Wanke, S. (2021). Discordant phylogenomic placement of Hydnoraceae and Lactoridaceae within Piperales using data from all three genomes. *Front. Plant. Sci.* 12:642598. doi: 10.3389/fpls.2021.642598

Junier, T., and Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the unix shell. *Bioinformatics* 26, 1669–1670. doi: 10.1093/bioinformatics/btq243

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kapli, P., Yang, Z., and Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* 21, 428–444. doi: 10.1038/s41576-020-0233-0

Katoh, K., Kuma, K.-I., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511–518. doi: 10.1093/nar/gki198

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kellogg, E., and Linder, H. (1995). "Phylogeny of Poales," in *Monocotyledons: Systematics and Evolution*, eds P.J. Rudall, P.J. Cribb, D.F. Cutler, and C.J. Humphries (London: Royal Botanic Gardens, Kew), 511-542.

Kubešová, M., Moravcová, L., Suda, J., Jarošík, V., and Pyšek, P. (2010). Naturalized plants have smaller genomes than their non-invading relatives: a flow cytometric analysis of the Czech alien flora. *Preslia* 82, 81–96.

Larson, D. A., Walker, J. F., Vargas, O. M., and Smith, S. A. (2020). A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. *Am. J. Bot.* 107, 773–789. doi: 10.1002/ajb2.1469

Lee, C., Ruhlman, T. A., and Jansen, R. K. (2020). Unprecedented intraindividual structural heteroplasmy in eleocharis (Cyperaceae. Poales) Plastomes. *Genome Biol. Evol.* 12, 641–655. doi: 10.1093/gbe/evaa076

Lee, J. M., Song, H. J., Park, S. I., Lee, Y. M., Jeong, S. Y., Cho, T. O., et al. (2018). Mitochondrial and plastid genomes from coralline red algae provide insights into the incongruent evolutionary histories of organelles. *Genome Biol. Evol.* 10, 2961–2972. doi: 10.1093/gbe/evy222

Leseberg, C. H., and Duvall, M. R. (2009). The complete chloroplast genome of Coix lacryma-jobi and a comparative molecular evolutionary analysis of plastomes in Cereals. *J. Mol. Evol.* 69, 311–318. doi: 10.1007/s00239-009-9275-9

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., et al. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* 5, 461–470. doi: 10.1038/s41477-019-0421-0

Linder, H. P., and Rudall, P. J. (1993). The megagametophyte in anarthria (Anarthriaceae, Poales) and its implications for the phylogeny of the Poales. *Am. J. Bot.* 80, 1455–1464. doi: 10.2307/2445675

Linder, H. P., and Rudall, P. J. (2005). Evolutionary history of Poales. *Annu. Rev. Ecol. Evol. Syst.* 36, 107–124.

Ma, P. F., Vorontsova, M. S., Nanjarisoa, O. P., Razanatsoa, J., Guo, Z. H., Haevermans, T., et al. (2017). Negative correlation between rates of molecular evolution and flowering cycles in temperate woody bamboos revealed by plastid phylogenomics. *BMC Plant Biol.* 17:260. doi: 10.1186/s12870-017-1199-8

Ma, P. F., Zhang, Y. X., Zeng, C. X., Guo, Z. H., and Li, D. Z. (2014). Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (poaceae). *Syst. Biol.* 63, 933–950. doi: 10.1093/sysbio/syu054

McKain, M. R., Tang, H., McNeal, J. R., Ayyampalayam, S., Davis, J. I., et al. (2016). A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biol. Evol.* 8, 1150–1164. doi: 10.1093/gbe/evw060

Michelangeli, F. A., Davis, J. I., and Stevenson, D. W. (2003). Phylogenetic relationships among Poaceae and related families as inferred from morphology, inversions in the plastid genome, and sequence data from the mitochondrial and plastid genomes. *Am. J. Bot.* 90, 93–106. doi: 10.3732/ajb.90.1.93

Morales-Briones, D. F., Liston, A., and Tank, D. C. (2018). Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus Lachemilla (Rosaceae). *New Phytol.* 218, 1668–1684. doi: 10.1111/nph.15099

Mu, X.-Y., Tong, L., Sun, M., Zhu, Y.-X., Wen, J., Lin, Q.-W., et al. (2020). Phylogeny and divergence time estimation of the walnut family (Juglandaceae) based on nuclear RAD-Seq and chloroplast genome data. *Mol. Phylogenet. Evol.* 147:106802. doi: 10.1016/j.ympev.2020.106802

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Norman, J. E., and Gray, M. W. (2001). A complex organization of the gene encoding cytochrome oxidase subunit 1 in the mitochondrial genome of the dinoflagellate, crypthecodinium cohnii: homologous recombination generates two different cox1 open reading frames. *J. Mol. Evol.* 53, 351–363. doi: 10.1007/s002390010225

Parks, M., Cronn, R., and Liston, A. (2012). Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus L.* (Pinaceae). *BMC Evol. Biol.* 12:100. doi: 10.1186/1471-2148-12-100

Pease, J. B., Brown, J. W., Walker, J. F., Hinchliff, C. E., and Smith, S. A. (2018). Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. *Am. J. Bot.* 105, 385–403. doi: 10.1002/ajb2.1016

Peredo, E. L., King, U. M., and Les, D. H. (2013). The plastid genome of *Najas flexilis*: adaptation to submersed environments is accompanied by the complete loss of the NDH complex in an aquatic angiosperm. *PLoS One* 8:e68591. doi: 10.1371/journal.pone.0068591

Perrotta, G., Grienenberger, J. M., and Gualberto, J. M. (2002). Plant mitochondrial rps2 genes code for proteins with a C-terminal extension that is processed. *Plant Mol. Biol.* 50, 523–533. doi: 10.1023/a:1019878212696

Qiu, Y.-L., Li, L., Wang, B., Xue, J.-Y., Hendry, T. A., Li, R.-Q., et al. (2010). Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *J. Syst. Evol.* 48, 391–425. doi: 10.1111/j.1759-6831.2010.00097.x

Qu, X. J., Moore, M. J., Li, D. Z., and Yi, T. S. (2019). PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* 15:50. doi: 10.1186/s13007-019-0435-7

Rambaut, A. (2012). Figtree v1.4.4. https://github.com/rambaut/figtree/releases/tag/v1.4.4 (accessed November 26, 2018).

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B. F., and Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* 56, 389–399. doi: 10.1080/10635150701397643

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Saarela, J. M., Burke, S. V., Wysocki, W. P., Barrett, M. D., Clark, L. G., Craine, J. M., et al. (2018). A 250 plastome phylogeny of the grass family (Poaceae): topological support under different data partitions. *PeerJ* 6:e4299. doi: 10.7717/peerj.4299

Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. Evol.* 1:0126. doi: 10.1038/s41559-017-0126

Shen, X.-X., Steenwyk, J. L., and Rokas, A. (2021). Dissecting incongruence between concatenation- and quartet-based approaches in phylogenomic data. *Syst. Biol.* 70, 997–1014. doi: 10.1093/sysbio/syab011

Šmarda, P., Bureš, P., Horová, L., Leitch, I. J., Mucina, L., Pacini, E., et al. (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc. Natl. Acad. Sci. U S A.* 111, E4096–E4102. doi: 10.1073/pnas.1321152111

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* 15:150. doi: 10.1186/s12862-015-0423-0

Soltis, D. E., Smith, S. A., Cellinese, N., Wurdack, K. J., Tank, D. C., Brockington, S. F., et al. (2011). Angiosperm phylogeny: 17 genes, 640 taxa. *Am. J. Bot.* 98, 704–730. doi: 10.3732/ajb.1000404

Soltis, P. S., and Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annu. Rev. Plant. Biol.* 60, 561–588. doi: 10.1146/annurev.arplant.043008.092039

Som, A. (2015). Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.* 16, 536–548. doi: 10.1093/bib/bbu015

Stamatakis, A. (2015). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stegemann, S., Keuthe, M., Greiner, S., and Bock, R. (2012). Horizontal transfer of chloroplast genomes between plant species. *Proc. Natl. Acad. Sci. U S A.* 109, 2434–2438. doi: 10.1073/pnas.1114076109

Stevens, P. F. (2001). *Angiosperm Phylogeny Website, Version 14*. Available online at: http://www.mobot.org/MOBOT/Research/APweb/ (accessed May, 22, 2021)

Sun, M., Soltis, D. E., Soltis, P. S., Zhu, X., Burleigh, J. G., and Chen, Z. (2015). Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Mol. Phylogenet. Evol.* 83, 156–166. doi: 10.1016/j.ympev.2014.11.003

Sun, X. Z. (1992). "Typhaceae," in *Flora of China Section 13*, ed. Flora of China Editorial Committee (Beijing: Science Press).

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Talavera, G., Lukhtanov, V., Pierce, N. E., and Vila, R. (2021). DNA barcodes combined with multi-locus data of representative taxa can generate reliable higher-level phylogenies. *Syst. Biol.* syab038. doi: 10.1093/sysbio/syab038 [Epub ahead of print].

Tian, X., and Li, D. Z. (2002). Application of DNA sequences in plant phylogenetic study. *Acta Bot. Yunnanica* 24, 170–184.

Triplett, J. K., Oltrogge, K. A., and Clark, L. G. (2010). Phylogenetic relationships and natural hybridization among the North American woody bamboos

(Poaceae: Bambusoideae: Arundinaria). *Am. J. Bot.* 97, 471–492. doi: 10.3732/ajb.0900244

Vallée, G. C., Muñoz, D. S., and Sankoff, D. (2016). Economic importance, taxonomic representation and scientific priority as drivers of genome sequencing projects. *BMC Genomics.* 17, 782. doi: 10.1186/s12864-016-3100-9

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424. doi: 10.1038/nrg.2017.26

Vargas, O. M., Ortiz, E. M., and Simpson, B. B. (2017). Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: Diplostephium). *New Phytol.* 214, 1736–1750. doi: 10.1111/nph.14530

Walker, J. F., Walker-Hale, N., Vargas, O. M., Larson, D. A., and Stull, G. W. (2019). Characterizing gene tree conflict in plastome-inferred phylogenies. *PeerJ* 7:e7747. doi: 10.1101/512079

Wei, W., Youliang, Z., Li, C., Yuming, W., Zehong, Y., and Ruiwu, Y. (2005). PCR-RFLP analysis of cpDNA and mtDNA in the genus Houttuynia in some areas of China. *Hereditas* 142, 24–32. doi: 10.1111/j.1601-5223.2005.01704.x

Wen, J., Liu, J., Ge, S., Xiang, Q.-Y., and Zimmer, E. A. (2015). Phylogenomic approaches to deciphering the tree of life. *J. Syst. Evol.* 53, 369–370. doi: 10.1111/jse.12175

Wendel, J. F., Schnabel, A., and Seelanan, T. (1995). An unusual ribosomal DNA sequence from gossypium gossypioides reveals ancient, cryptic, intergenomic introgression. *Mol. Phylogenet. Evol.* 4, 298–313. doi: 10.1006/mpev.1995.1027

Wu, S., Han, B., and Jiao, Y. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol. Plant.* 13, 59–71. doi: 10.1016/j.molp.2019.10.012

Xi, Z., Ruhfel, B. R., Schaefer, H., Amorim, A. M., Sugumaran, M., Wurdack, K. J., et al. (2012). Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc. Natl. Acad. Sci. U S A.* 109, 17519–17524. doi: 10.1073/pnas.1205818109

Yang, Y.-Y., Qu, X.-J., Zhang, R., Stull, G. W., and Yi, T.-S. (2021). Plastid phylogenomic analyses of Fagales reveal signatures of conflict and ancient chloroplast capture. *Mol. Phylogenet. Evol.* 163:107232. doi: 10.1016/j.ympev.2021.107232

Yu, L., and Zhang, Y. P. (2006). Phylogenomics–an attractive avenue to reconstruct "tree of life". *Hereditas* 28, 1445–1450. doi: 10.1360/yc-006-1445

Zeng, L., Zhang, N., and Ma, H. (2014). Advances and challenges in resolving the angiosperm phylogeny. *Biodiversity Sci.* 22:21. doi: 10.3724/SP.J.1003.2014.13189

Zeng, L., Zhang, N., Zhang, Q., Endress, P. K., Huang, J., and Ma, H. (2017). Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytol.* 214, 1338–1354. doi: 10.1111/nph.14503

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

Zhang, D., Gao, F., Jakovliæ, I., Zou, H., Zhang, J., Li, W. X., et al. (2020). PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour.* 20, 348–355. doi: 10.1111/1755-0998.13096

Zhang, R., Wang, Y.-H., Jin, J.-J., Stull, G. W., Bruneau, A., Cardoso, D., et al. (2020). Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. *Syst. Biol.* 69, 613–622. doi: 10.1093/sysbio/syaa013

Zhang, S.-D., Jin, J.-J., Chen, S.-Y., Chase, M. W., Soltis, D. E., Li, H.-T., et al. (2017). Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. *New Phytol.* 214, 1355–1367. doi: 10.1111/nph.14461

Zhao, F., Chen, Y. P., Salmaki, Y., Drew, B. T., Wilson, T. C., Scheen, A. C., et al. (2021). An updated tribal classification of Lamiaceae based on plastome phylogenomics. *BMC Biol.* 19:2. doi: 10.1186/s12915-020-00931-z

Zou, X. H., and Ge, S. (2008). Conflicting gene trees and phylogenomics. *J. Syst. Evol.* 46:795. doi: 10.3724/SP.J.1002.2008.08081

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 24 March 2022
doi: 10.3389/fpls.2022.811206

Cryptic Species Diversification of the *Pedicularis siphonantha* Complex (Orobanchaceae) in the Mountains of Southwest China Since the Pliocene

*Rong Liu[1,2,3], Hong Wang[4], Jun-Bo Yang[5], Richard T. Corlett[1,2], Christopher P. Randle[6], De-Zhu Li[5] and Wen-Bin Yu[1,2,7]**

[1] Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, China, [2] Center of Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences, Mengla, China, [3] University of Chinese Academy of Sciences, Beijing, China, [4] Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China, [5] Plant Germplasm and Genomics Centre, Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China, [6] Department of Biological Sciences, Sam Houston State University, Huntsville, TX, United States, [7] Southeast Asia Biodiversity Research Institute, Chinese Academy of Sciences, Yezin, Myanmar

OPEN ACCESS

Edited by:
Susann Wicke, Humboldt University of Berlin, Germany

Reviewed by:
Richard Ree, Field Museum of Natural History, United States
Chun-Lei Xiang, Kunming Institute of Botany (CAS), China
Alex Twyford, University of Edinburgh, United Kingdom

***Correspondence:**
Wen-Bin Yu
yuwenbin@xtbg.ac.cn

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *08 November 2021*
Accepted: *21 February 2022*
Published: *24 March 2022*

Citation:
Liu R, Wang H, Yang J-B, Corlett RT, Randle CP, Li D-Z and Yu W-B (2022) Cryptic Species Diversification of the Pedicularis siphonantha Complex (Orobanchaceae) in the Mountains of Southwest China Since the Pliocene. Front. Plant Sci. 13:811206. doi: 10.3389/fpls.2022.811206

Morphological approaches often fail to delimit species in recently derived species complexes. This can be exacerbated in historical collections which may have lost key features in specimen preparation and preservation. Here, we examine the *Pedicularis siphonantha* complex, endemic to the Mountains of Southwest China. This complex is characterized by its red/purple/pink and long-tubular corolla, and twisted, beaked galea. However, herbarium specimens are often difficult to identify to species. Molecular approaches using nrITS or nuclear ribosomal internal transcribed spacer (nrITS) + plastid DNA (ptDNA) have been successfully used for species identification in *Pedicularis*. To resolve taxonomic confusion in the *Pedicularis siphonantha* complex, we reconstructed phylogenies of the complex using nrITS and four plastid DNA loci (*matK*, *rbcL*, *trnH-psbA*, and *trnL-F*). To recover as much of the phylogenetic history as possible, we sampled individuals at the population level. Topological incongruence between the nrITS and ptDNA datasets was recovered in clades including two widely distributed species, *Pedicularis milliana* and *Pedicularis tenuituba*. Based on morphological, geographical, and genetic evidence, we suggest that hybridization/introgression has occurred between *P. milliana* and *Pedicularis sigmoidea*/*Pedicularis* sp. 1 in the Yulong Snow Mountain of Lijiang, northwest Yunnan, and between *P. tenuituba* and *Pedicularis leptosiphon* in Ninglang, northwest Yunnan. After removing conflicting DNA regions in *Pedicularis dolichosiphon* (nrITS) and *P. milliana* (ptDNA), the concatenated nrITS and ptDNA phylogenies distinguish 11 species in the *P. siphonantha* complex, including two undescribed species, from the Jiaozi and Yulong Snow Mountains, respectively. Phylogeographical analyses indicate that the *P. siponantha* complex originated from south of the Hengduan Mountains, expanding north to the Himalayas and the Yunnan-Guizhou Plateau. Moreover, the uplift of the Qinghai-Tibet Plateau and climate oscillations may have driven further diversification in the complex.

Keywords: *Pedicularis siphonantha* complex, phylogenetic delimitation, speciation, mountains of Southwest China, the Hengduan Mountains

INTRODUCTION

The Mountains of Southwest China host one of the richest temperate floras, with a high proportion of endemic species (Boufford, 2014). Mountain uplifts and the monsoon climate create geographically and ecologically isolated habitats, which have driven plant diversification in this region (Hoorn et al., 2013; Xing and Ree, 2017; Ding et al., 2020). Rapid diversification and frequent introgression compound taxonomic confusion, as documented in megadiverse genera of the region, such as *Meconopsis* Vig. (Papaveraceae), *Primula* L. (Primulaceae), and *Rhododendron* L. (Ericaceae) (Zha et al., 2010; Yang et al., 2012; Favre et al., 2016). Species delimitation is traditionally based on morphological characters. However, morphological approaches often fail to delimit recently diverged species, resulting in cryptic species complexes (Bickford et al., 2007; Struck et al., 2018). The study of cryptic species, therefore, offers a window into the diversification and the maintenance of recently divergent species groups.

Pedicularis L. (Orobanchaceae) consists of approximately 600–800 species, of which two-thirds are endemic to the Mountains of Southwest China (Li, 1948, 1949). *Pedicularis* exhibits dramatic variations in corolla structure and galea form, beak length and shape, and corolla tube length, which are key characters for species delimitation. Four general corolla types are recognized: (A) short-tubular corolla with a beakless, toothless galea (upper lip), (B) short-tubular corolla with a toothed galea, (C) short-tubular corolla with a beaked galea, and (D) long-tubular corolla with a beaked galea (Maximowicz, 1888; Li, 1948, 1949; Tsoong, 1955, 1956, 1963). Long-tubular corollas always bear a beaked galea, which has been considered as a derived corolla type. Ree (2005) and Yu et al. (2015) have demonstrated that long-tubular corollas appear to have been derived from short-tubular corollas several times, resulting in taxonomic confusion. The *Pedicularis siphonantha* complex includes only long-tubular species with the purple-red corolla and an S-shaped, beaked galea (**Figure 1**). *Pedicularis siphonantha* D. Don from the Himalayas was the first described species. To date, at least 11 species of this complex are recognized from the western Himalayas to the Mountains of Southwest China (Li, 1949; Yang et al., 1998; Yu et al., 2015, 2018). The *P. siphonantha* complex was supported as monophyletic by Yu et al. (2015). However, species within this complex are difficult to distinguish, especially as herbarium specimens, which lose three-dimensional corolla structure and color and often lack field photos and descriptions of key diagnostic characters. This results in uncertainty about the number of species and their geographical distributions. For example, pressed specimens of *Pedicularis delavayi* Franch. ex Maxim. resemble *P. siphonantha*, though the three-dimensional structure of the middle lobe of the lower lip and corolla throat color allows easy distinction. Because of the loss of structural and color characteristics in preserved specimens, Tsoong (1963) inferred that the distribution of *P. siphonantha* extended to the Mountains of Southwest China, and included the species as a variety of *P. siphonantha*, i.e., *P. siphonantha* var. *delavayi* (Franch. ex Maxim.) P. C. Tsoong. In contrast, Li (1949) considered *P. delavayi* as a separated species. Li (1949) and Tsoong (1963), as well as Yang et al. (1998) and other botanists, misidentified most herbarium specimens of *Pedicularis tenuituba* H. L. Li and *Pedicularis milliana* W. B. Yu et al. as *P. delavayi* (Yu et al., 2018). In addition, the infraspecific taxonomy of *P. siphonantha* is not fully resolved yet in the Himalaya region (Yu et al., 2018).

Molecular approaches have been widely applied for species identification (Hebert et al., 2003; Hebert and Gregory, 2005; Hollingsworth et al., 2016; Kress, 2017). Four candidate DNA barcodes [nuclear ribosomal internal transcribed spacer (nrITS), and three plastid *matK*, *rbcL*, and *trnH-psbA* regions] can be used to discriminate more than 89.0% of *Pedicularis* species (Yu et al., 2011; Liu et al., 2013). Based on nrITS and four plastid loci (*matK*, *rbcL*, *trnH-psbA,* and *trnL-F*), no samples of *P. delavayi* cluster with other species of the *P. siphonanta* complex. Yu et al. (2018) have reinstated *P. delavayi* as a separate species and discovered an undescribed species, *P. milliana,* which was previously misidentified as *P. siphonantha* var. *delavayi* or *P. delavayi*. To date, species delimitation of the *P. siphonantha* complex is not fully resolved, due to limited sampling and poorly known species distributions (Li, 1949; Yu et al., 2015, 2018). For example, though *P. milliana* occurs widely in northwest Yunnan and *Pedicularis sigmoidea* Franch. ex Maxim. occurs in the south margin of *P. milliana* (**Figure 2**), specimens (see **Figure 1G**) from the Yulong Snow Mountain (G1 in **Figure 2**) bear a remarkable S-shaped beak similar to that of *P. sigmoidea*. Other specimens (see **Figure 1E**), from the Jiaozi Snow Mountain (E1-3 in **Figure 2**), appear to be distinct from known species (Yu et al., 2015, 2018). Morphological ambiguity is supported by topological incongruence between nrITS and plastid gene datasets (Yu et al., 2015), which might be caused by hybridization or introgression.

In this study, we reconstructed a comprehensive phylogeny of the *P. siphonantha* complex using five DNA loci (nrITS, *matK, rbcL, trnH-psbA*, and *trnL-F*) with population-level sampling. Our main goals were to: (1) explore patterns and causes of phylogenetic incongruence between nrITS and plastid DNA datasets in the *P. siphonantha* complex; (2) revise species delimitations in the *P. siphonantha* complex; and (3) investigate the causes of species diversification in the *P. siphonantha* complex.

MATERIALS AND METHODS

Taxon Sampling

We sampled 78 individuals, mainly from the Hengduan Mountains, as well as the Himalayas and the Yunnan-Guizhou Plateau, representing 11 taxa of the *P. siphonantha* complex and 12 other *Pedicularis* species (**Supplementary Table 1**). The 11 taxa covered nine recognized species of the complex, with the exception of *Pedicularis fastigiata* Franch., which is known only from the type collection by Orléans H.d' s.n. (P, barcode P00520823) of the 78 individuals, 50 were newly sampled and sequenced. There were 22 samples (19 populations) of *P. tenuituba*, widely distributed in western Sichuan, and 17 samples (14 populations) of *P. milliana*, endemic to the

FIGURE 1 | Field photos of 11 species of the *Pedicularis siphonantha* complex and *Pedicularis delavayi*, **(A)** *Pedicularis tenuituba* H. L. Li; **(B)** *Pedicularis leptosiphon* H. L. Li; **(C)** *Pedicularis dolichosiphon* (Hand.-Mazz.) H. L. Li; **(D)** *P. siphonantha* D. Don; **(E)** *Pedicularis* sp. 2 from Jiaozi snow mountain; **(F)** *Pedicularis milliana* W. B. Yu, D. Z. Li, and H. Wang; **(G)** *Pedicularis* sp. 1 from Ganheba, Lijiang; **(H)** *Pedicularis sigmoidea* Franch. ex Maxim; **(I)** *Pedicularis variegata* H. L. Li; **(J)** *Pedicularis dolichantha* Bonati; **(K)** *Pedicularis humilis* Bonati. **(L)** *Pedicularis delavayi* Franch. ex Maxim.

northwestern Yunnan. Three populations of *P. milliana* (i.e., F11, F12, and F13) were collected from the Yulong Snow Mountain in Lijiang, and population F14 was collected from the Haba Snow Mountain in Shangri-La. In addition, population G (i.e., sample LIDZ1584), collected from Ganheba in the Yulong Snow Mountain represents an unknown taxon, which is similar to *P. sigmoidea* in the shape of the galea beak but has a smaller corolla. Populations E1–E3, collected from the Jiaozi Snow Mountain, represents another unknown taxon, which is distinguished from *P. milliana* by its oblate and crested beak. The remaining six taxa of the *P. siphonantha* complex have narrow distributions, so only a few individuals/populations were included in this study. We included nine samples (seven populations) of *P. delavayi* from the northwestern Yunnan and western and northern Sichuan, where it overlaps with *P. milliana* and *P. tenuituba*. Geographic information for all samples the *P. siphonantha* complex is shown in **Figure 2**.

Molecular Methods

The nrITS and four plastid DNA (ptDNA) markers (*matK*, *rbcL*, *trnH-psbA*, and *trnL-F*) were amplified and sequenced in this study. Primer information of five DNA markers were presented in previous studies (Yu et al., 2011, 2013, 2018). Genomic DNA of 50 new samples was extracted using a modified CTAB method from silica gel–dried leaves. PCR amplification and sequencing profile followed Yu et al. (2011). Raw sequences were assembled and edited using Geneious 7.1 (Kearse et al., 2012). The nrITS region has multiple copies in the genome. These copies showed evolutionary consistent in the newly sequenced 46 samples, only one sample (*P. milliana* F2/03-060) had two ambiguous basecalls (i.e., multiple superimposed peaks in chromatograms), and three samples (*P. milliana* F2/03-059, *P. tenuituba* A11/HW10187, and *P. tenuituba* A7/HW10327) had one basecall. The ambiguous site was assigned using IUPAC ambiguity characters. Assembled sequences were aligned using MAFFT 7.4 (Katoh et al., 2019), then adjusted manually using Geneious. Sequence characteristics and Kimura 2-parameter (K-2P) model-based genetic distances among taxa were calculated using MEGA 10.0 (Kumar et al., 2018), and non-parametric two-sample Kolmogorov–Smirnov (K–S) tests of genetic distances between and within species were estimated for widely distributed species *P. tenuituba*, *P. milliana*, and *P. siphonantha* using SPSS 25.0 (IBM, 2017). The nrITS and ptDNA datasets were analyzed separately.

Phylogenetic Analyses

Both Bayesian Inference (BI) and maximum likelihood (ML) were used to reconstruct phylogenetic relationships in the *P. siphonantha* complex. To explore topological incongruence

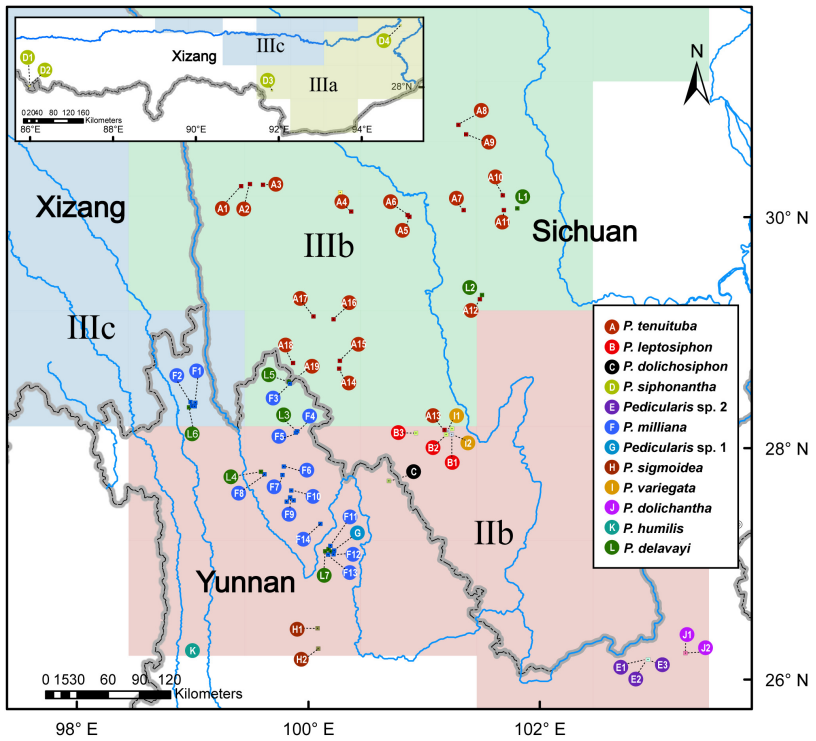


FIGURE 2 | A geographical sampling of the *Pedicularis siphonantha* complex from the Hengduan Mountains, southwestern China. The colors of circles correspond to different taxa. Numbers and letters indicate populations; these abbreviations are also used in **Figures 3**, **4**, **6**. More information regarding collection vouchers can be found in **Supplementary Table 1**.

TABLE 1 | The best-fit model of partition dataset partitions.

DNA marker	nrITS	ptDNA			
		matK	*rbcL*	*trnH-psbA*	*trnL-F*
BIC model	GTR+G4	HKY+G4	K80+I	F81+G4	GTR+I+G4
-lnL	2248.9747	1785.4927	1244.1168	1854.9583	2345.2173
K	9	5	2	4	10
Frequency A	0.1973	0.3577	0.2500	0.3973	0.3678
Frequency C	0.2946	01791	0.2500	0.1061	0.1663
Frequency G	0.2802	0.1725	0.2500	0.1037	0.1527
A↔C	0.9500	1.0000	1.0000	1.0000	1.7725
A↔G	0.9996	3.5998	4.0355	1.0000	1.4874
A↔T	1.3845	1.0000	1.0000	1.0000	0.2393
C↔G	0.2899	1.0000	1.0000	1.0000	0.5144
C↔T	3.5963	3.5998	4.0355	1.0000	1.5535
G↔T	1.0000	1.0000	1.0000	1.0000	1.0000
Gamma distribution shape parameter of variable sites	0.2744	0.2797	0.0000	0.2844	0.5634
Proportion of invariable sites	0.0000	0.0000	0.8783	0.0000	0.6024

between nrITS and the concatenated ptDNA phylogenies, the two datasets were analyzed separately. Before concatenating the data sets, we removed sequences that seemed to be the source of conflict. The concatenated datasets (ptDNA and nrITS + ptDNA) were partitioned by gene and the best-fit model was estimated using Modeltest-ng (Darriba et al., 2020) (see **Table 1**). BI Markov chain Monte Carlo (MCMC) analyses were performed using MrBayes 3.2 (Ronquist et al., 2012) for 2,000,000 generations with two simultaneous runs, each comprising four incrementally heated chains. BI analyses were started with random trees and sampled every 1,000 generations. The first 25% of trees were discarded as burn-in, and the remaining trees were used to generate a majority-rule consensus tree. Posterior probability (PP) values ≥0.95 were considered as well-supported

(Alfaro et al., 2003; Kolaczkowski and Thornton, 2007). ML tree search was performed using RAxML version 8.2.12 (Stamatakis, 2014) under GTR + Γ. Node support was evaluated using 1,000 non-parametric bootstrap (BS) replicates. Nodes with BS values ≥70 were considered well-supported (Hillis and Bull, 1993).

To explore patterns of introgression in the complex, we constructed phylogenetic networks of the *P. siphonantha* complex based on the total dataset by the concatenation of all nrITS and ptDNA sequences by using SplitsTree 4.14.1 (Huson and Bryant, 2006). The Neighbor-net model was performed using the Kimura 2-parameter (K-2P) distance and Ordinary Least Squares Method, with 1,000 BS replicates to estimate split support. Splits with BS ≥ 70 were considered as well-supported.

Topological Conflict Analyses

Thresholds of PP ≥ 0.95 and BS ≥ 70 were interpreted as identifying incongruent clades between the nrITS and ptDNA datasets. Based on topological incongruence between the nrITS and ptDNA datasets, the DNA sequence would be considered as heterogeneous one if the phylogenetic cluster was not consistent with the morphological cluster, then the heterogeneous sequence was removed from the concatenated nrITS + ptDNA dataset. Herein, the nrITS sequence of *Pedicularis dolichosiphon* (Hand.-Mazz.) H. L. Li and the four ptDNA regions (*matK*, *rbcL*, *trnH-psbA*, and *trnL-F*) of the samples F11–F14 of *P. milliana* were identified as heterogeneous sequences, so that those sequences were removed from the concatenated nrITS + ptDNA dataset. Then, the concatenated nrITS + ptDNA phylogeny was performed using the same methods as nrITS and the concatenated ptDNA phylogenetic analyses (see above). Additionally, the Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) and the approximately unbiased (AU) test (Shimodaira, 2002) were used to estimate the degree of topological incongruence among the three datasets (nrITS, ptDNA, and modified nrITS + ptDNA). Constraint trees were constructed in Mesquite version 3.6 (Maddison and Maddison, 2019), and the SH and AU tests were performed using IQ-Tree 1.6 (Lam-Tung et al., 2015).

Phylogeographical Analyses

Ancestral geographical distributions were inferred using BEAST 2.6.3 (Bouckaert et al., 2014). The concatenated nrITS and ptDNA dataset, including all samples of the *P. siphonantha* complex and two outgroups *Pedicularis amplituba* H. L. Li and *Pedicularis tachanensis* Bonati, was imported into BEAUti with "beast-classic package" (Lemey et al., 2009). Samples were assigned to one of four subregions of the Sino-Himalayan flora following Liu et al. (2021). Taking into account the center of endemism identified by Zhang et al. (2016), two samples of *Pedicularis siphonantha* (D1, D2) collected from Yadong country in the middle Himalaya were assigned to subregion IIIa. The sites model was calibrated using the "bModeltest package" (Bouckaert and Drummond, 2017), the molecular clock model was set to "Relaxed Clock Log Normal," and the Yule model served as the tree prior. The divergence time of the most recent ancestor between the *P. siphonantha* complex and the two outgroups was constrained to 9.1 ± 2 Mya. A second calibration point was obtained from the analysis of Yu et al. (2015) (**Supplementary Figure 1**). MCMC chains were run for 10,000,000 generations, with parameter values and trees sampled every 1,000 generations. Effective sample size (ESS > 200) was assessed using Tracer 1.7 (Rambaut et al., 2018). After discarding 25% of the initial trees as burn-in, the maximum clade credibility (MCC) tree with mean ages and 95% highest posterior density (HPD) intervals on nodes was reconstructed using TreeAnnotator 2.6.3 (Bouckaert et al., 2014).

RESULTS

Matrix Characteristics

Matrix characteristics of nrITS, four plastid DNA, and the concatenated ptDNA datasets are shown in **Table 2**. Similar to previous studies (Yu et al., 2011, 2013, 2015, 2018), these five loci included an adequate numbers of variable sites and parsimony-informative sites for subsequent phylogenetic analyses in the *P. siphonantha* complex. The nrITS dataset is the most informative, followed by two plastid intergenic spacer datasets (*trnH-psbA* and *trnL-F*). The two protein-coding genes (*rbcL* and *matK*) are less informative than the three spacer datasets. Sequences of *P. siphonantha* and *P. milliana* show the highest variation at the species level for both nrITS and ptDNA datasets.

Genetic Distance Estimation

Comparisons of genetic distances within and between species using the K–S test are shown in **Figure 3**. The intraspecific genetic distances among *P. milliana* sequences were significantly smaller ($P < 0.05$) than the distance between *P. milliana* and any other taxon, with one exception; nrITS sequences of *P. milliana* and *Pedicularis* sp. 1 did not differ significantly. The same pattern holds for *P. tenuituba*, with the exception being a non-significant distance with nrITS sequences *P. dolichosiphon*. While the ptDNA distances within *P. siphonantha* were significantly less than the distance between *P. siphonantha* and ptDNA of any other taxon, distances of nrITS sequences within *P. siphonantha* were significantly smaller than the distance between *P. milliana* and *Pedicularis tahaiensis* Bonati ($P < 0.05$).

Phylogenetic Analyses of the nrITS Dataset

Maximum likelihood and BI analyses obtained identical topologies for the nrITS dataset (**Figure 4A**). The *P. siphonantha* complex was recovered as monophyletic (BS/PP = 95/1.00), and *P. delavayi* was not included in that clade. Within the *P. siphonantha* complex, five major clades were recovered. Clade I had only *Pedicularis humilis* Bonati, and it was moderately supported as sister to the remaining four clades (BS/PP = 63/0.95). The relationships among Clades II–V were not well-resolved. Clade II consisted of three individuals of *Pedicularis* sp. 2, and clade III included two species, *Pedicularis leptosiphon* H. L. Li and *P. siphonantha*, with *P. leptosiphon* nested within *P. siphonantha*, although with weak support. Clades IV and V were weakly supported as sister lineages (BS/PP = 55/0.87). Clade IV consisted of five species, *Pedicularis dolichantha* Bonati, *P. milliana*, *P. sigmoidea, Pedicularis variegata* H. L. Li, and

TABLE 2 | Sequence characteristics of nrITS and four plastid DNA regions.

Parameters	*n*	nrITS	Plastid DNA loci				Concatenated datasets	Total
			matK	*rbcL*	*trnH-psbA*	*trnL-F*		
No. of accessions		77	77	77	71	76	78	78
Aligned length(bp)		723	856	727	725	987	3295	4,018
Variable sites/Parsimony informative sites								
P. siphonantha complex + Outgroups	78	150/103	127/70	43/31	128/99	135/89	417/388	566/490
P. delavayi	9	5/1	1/0	3/2	9/3	13/3	37/7	42/8
P. siphonantha complex	57	61/39	84/40	33/21	86/72	73/47	259/179	319/217
P. tenuituba	22	7/3	13/5	9/2	9/4	10/7	40/16	59/19
P. milliana	17	8/4	14/10	10/6	22/22	15/13	64/52	70/55
P. leptosiphon	2	1/0	0/0	3/0	0/0	1/0	4/0	5/0
P. siphonantha	4	19/7	10/1	0/0	12/0	11/1	33/2	52/9
Pedicularis sp. 2	3	1/0	4/0	0/0	1/0	4/0	7/0	7/0
P. variegata	2	0/0	0/0	0/0	0/0	4/0	4/0	4/0
P. sigmoidea	2	6/0	3/0	0/0		0/0	7/0	13/0
P. dolichantha	2	0/0	0/0	0/0	0/0	3/0	3/0	3/0

Pedicularis sp. 1 (G). In this clade, *Pedicularis* sp. 1 (G) was nested within *P. milliana*. Clade V consisted of *P. dolichosiphon* nested within 22 samples of *P. tenuituba*.

Phylogenetic Analyses of the ptDNA Dataset

Maximum likelihood and BI obtained identical topologies from the ptDNA dataset (**Figure 4B**). Unlike the nrITS phylogenies, the *P. siphonantha* complex was not recovered as monophyletic, including *P. amplituba* + *P. tachanensis* not considered here to be members of the complex. The core species of the *P. siphonantha* complex were split into two major clades (i.e., Clades A and B + C), with Clade C (*P. amplituba* and *P. tachanensis*) weakly supported as sister to the remaining members of Clade B (BS/PP = 38/0.65). Clade A included four species, *P. dolichosiphon*, *P. leptosiphon*, *P. siphonantha*, and *P. tenuituba*, which were strongly supported as monophyletic (BS/PP $\geq$ 99/1.00). Within Clade A, *P. siphonantha* was sister to the remaining three species with *P. dolichosiphon* and *P. leptosiphon* forming a clade sister to *P. tenuituba*.

Clade B contained seven species, which were split into two subclades. One subclade included *P. dolichantha*, *P. sigmoidea*, *Pedicularis* sp. 1, and four samples of *P. milliana* from the Yunlong Snow Mountain and the Haba Snow Mountain. In this subclade, *P. dolichantha* was sister to the remaining species, with *P. sigmoidea* as the sister to a clade including *Pedicularis* sp. 1 nested within *P. milliana*. The other subclade included *P. humilis*, *P. variegata*, *Pedicularis* sp. 2, and 13 samples of *P. milliana*, which were all supported as monophyletic. Within this subclade, *P. variegata* was sister to the remaining species, with *P. humulis* arising as sister to a clade including *Pedicularis* sp. 2 + *P. milliana*.

Topological Conflicts Between the nrITS and ptDNA Phylogenies

NrITS and ptDNA phylogenies are incongruent in the placement of *P. milliana*, *P. tenuituba,* and their relatives. In the nrITS phylogeny, 17 samples of *P. milliana* were only monophyletic if they included *Pedicularis* sp. 1 (BS/BP = 85/1.00). In the ptDNA phylogeny, they were separated into two distant clades, i.e., four samples (F11–F14) formed a clade by including *Pedicularis* sp. 1 (BS/BP = 100/1.00) as sister to *P. sigmoidea*, and the remaining 13 samples were monophyletic (BS/BP = 96/1.00) as sister to *Pedicularis* sp. 2. Twenty-four samples of *P. tenuituba* are strongly supported as monophyletic (BS/BP = 99/1.00) in the ptDNA phylogeny, but they are made paraphyletic by the inclusion of *P. dolichosiphon* in the nrITS phylogeny. Therefore, the four ptDNA regions of *P. milliana* samples F11–F14 and the nrITS sequence of *P. dolichosiphon* were identified as heterogeneous sequences. In addition, the nrITS dataset supports monophyly of the *P. siphonantha* complex (BS/BP = 95/100), but the ptDNA dataset does not, by including two short-tubular species *P. amplituba* and *P. tachanensis*.

Results of the AU and SH test for alternative hypotheses are summarized in **Table 3**. If phylogeny is constrained by the ptDNA dataset, SH and AU tests rejected ($P < 0.01$) the best tree topology of the nrITS dataset and monophyly of *P. milliana*. These tests failed to reject the monophyly of *P. tenuituba* + *P. dolichosiphon* and monophyly of the *P. siphonantha* complex. Meanwhile, when constrained by the nrITS dataset, SH and AU tests failed to reject ($P > 0.05$) the monophyly of *P. tenuituba* + *P. leptosiphon* + *P. siphonantha*, and the monophyly of *P. milliana* + *Pedicularis* sp. 2, but rejected the best tree topology of the ptDNA dataset, and the short tubular *P. tachanensis* + *P. amplituba* clade sister to the Clade B. Moreover, for the modified nrITS + ptDNA dataset, the null hypothesis that short tubular *P. tachanensis* + *P. amplituba* is sister to Clade 2 was rejected ($P < 0.05$).

Phylogenetic Analyses of the Concatenated nrITS + ptDNA Dataset

After removing conflicting sequences, *P. dolichosiphon* (nrITS) and *P. milliana* (ptDNA regions of F11–F14), ML and BI analyses produced nearly the same topology (**Figure 5**). Within

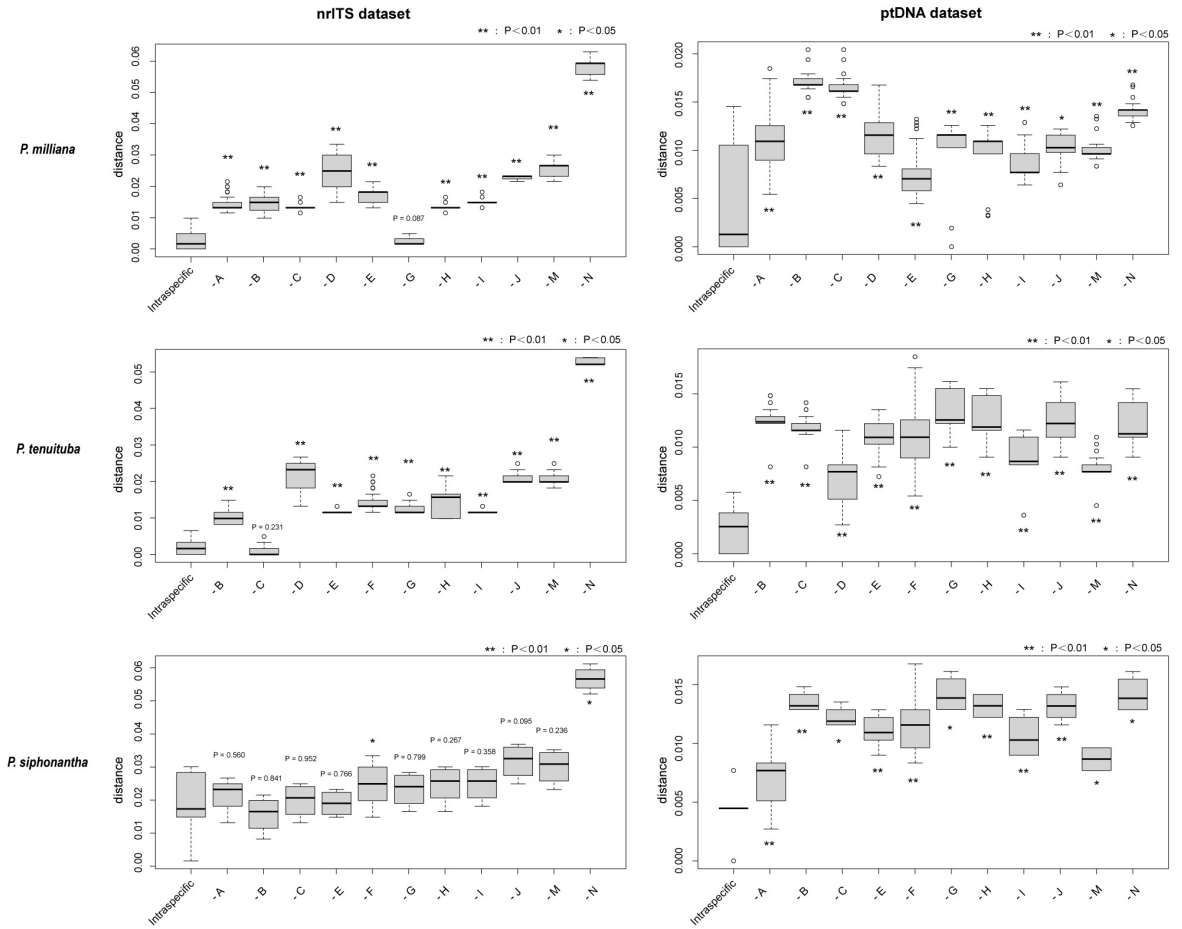


FIGURE 3 | K-S test for intraspecific and interspecific genetic distance. The intraspecific genetic distance of *Pedicularis milliana*, *Pedicularis tenuituba*, and *Pedicularis siphonantha* are estimated. Statistical significance was shown on every group (**$P < 0.01$, *$P < 0.05$). -A: to *Pedicularis tenuituba*; -B: to *Pedicularis leptosiphon*; -C: to *Pedicularis dolichosiphon*; -D: *to Pedicularis siphonantha*; -E: to *Pedicularis* sp. 2 from Jiaozi Snow Mountain; -F: to *Pedicularis milliana*; -G: to *Pedicularis* sp. 1 from Ganheba; -H: to *Pedicularis sigmoidea*; -I: to *Pedicularis variegata*; -J: to *Pedicularis dolichantha*; -M: to *Pedicularis amplituba* (outgroup); -N: to *Pedicularis tahaiensis* (outgroup).

the strongly monophyletic (BS/BP = 99/1.00) *P. siphonantha* complex, there were two major clades. Clade 1 (BS/BP = 96/1.00) included *P. dolichosiphon*, *P. leptosiphon*, *P. siphonantha*, and *P. tenuituba*. In this clade, *P. siphonantha* was sister to the remaining taxa, with the monophyletic *P. tenuituba* (BS/BP = 100/1.00) sister to a clade including *P. dolichosiphon* + *P. leptosiphon* (BS/BP = 100/1.00). Clade 2 (BS/BP = 95/1.00) included six taxa forming two subclades. In the larger subclade, *P. milliana* was monophyletic (BS/BP = 94/1.00), and sister to *Pedicularis* sp. 2 (BS/BP = 99/1.00). A clade including *P. milliana* + *Pedicularis* sp. 2 + *P. variegata* was sister to *P. humilis* (BS/BP = 65/0.93). In the other subclade, *P. dolichantha* was sister to *P. sigmoidea* + *Pedicularis* sp. 1 (BS/BP = 99/1.00).

Phylogenetic Network of the *Pedicularis siphonantha* Complex

The phylogenetic network of the concatenated nrITS and ptDNA dataset showed *P. milliana* split into two clusters (**Figure 6**), identical to those recovered in the ptDNA topology. Four samples from the Yulong Snow Mountain and the Haba Snow Mountain were nested with *Pedicularis* sp. 1 as sister to *P. sigmoidea* (BS = 96.8), and the other 13 samples were monophyletic as sister to *Pedicularis* sp. 2. In addition, *P. dolichosiphon* was resolved as either sister to *P. leptosiphon* (BS = 100) or *P. tenuituba* (BS = 85.6).

Phylogeographical Analyses of the *Pedicularis siphonantha* Complex

Phylogeographical analysis indicated that the most common ancestor of the *P. siphonantha* complex diverged from other *Pedicularis* in the late Miocene (6.04Mya–10.38Mya), south of the Hengduan Mountains (IIb), which harbors nine of eleven species/taxa of this complex (**Figure 7**). Species diversification of this complex mainly occurred in the Pliocene (2.48Mya–5.3Mya). After the initial divergence of

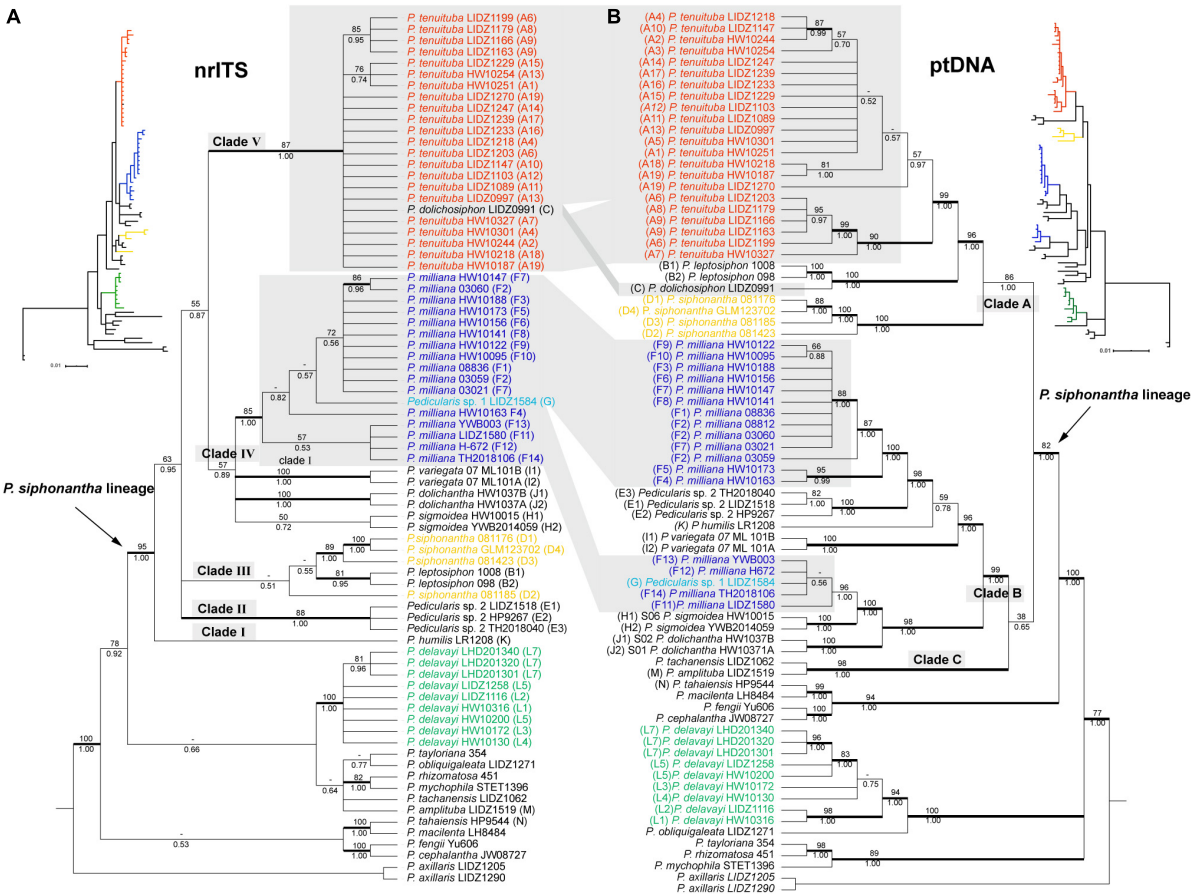


FIGURE 4 | Bayesian Inference (BI) tree comparisons of *Pedicularis siphonantha* complex inferred from nrITS **(A)** and the concatenated chloroplast **(B)** datasets. Phylogenetic incongruence is marked by shadow, Numbers associated with the branches are bootstrap value (BS) and BI posterior probabilities (PP), and thicker lines are indicated as BS ≥ 70 and PP ≥ 0.95. Node with BS < 50 was collapsed.

TABLE 3 | Summary of the Shimodaira–Hasegawa (SH) and the Approximately Unbiased (AU) tests.

Topological constraint	LnL	deltaL	p-AU	p-SH
ptDNA dataset	−7319.538			
Best tree of the nrITS phylogeny	−8004.596	685.060	<0.001	<0.001
Constraint the monophyly of *P. milliana*	−7412.726	93.188	<0.001	0.005
Constraint the clade *P. tenuituba* + *P. dolichosiphon*	−7319.538	0.000	0.545	1.000
Constraint the monophyly of the *P. siphonantha* complex	−7319.539	0.001	0.455	0.451
nrITS dataset	−2255.921			
Best tree of the ptDNA phylogeny	−2403.898	147.980	<0.001	<0.001
Constraint the monophyly of *P. tenuituba*	−2275.090	19.320	0.005	0.001
Constraint *P. tenuituba* + *P. leptosiphon* + *P. siphonantha*	−2255.921	0.000	0.988	1.000
Constraint *Pedicularis* sp. 2 (Jiaozi Mountain) sister to *P. milliana*	−2266.434	10.513	0.0147	0.569
Constraint *P. tachanensis* + *P. amplituba* sister to clade IV	−2282.702	26.782	0.002	0.248
Concanated nrITS + ptDNA dataset	−10765.132			
Constraint *P. tachanensis* + *P. amplituba* sister to clade 2	−10782.516	17.384	0.005	0.049

deltaL, logL difference from the maximal logL in the set.

Clades 1 and 2, members of Clade 1 migrated to the north and west, with Clade 2 diversifying *in situ*. In the Clade 1, *P. siphonantha* diverged and diversified in the Himalayan region (IIIa), *P. leptosiphon* and *P. dolichosiphon* diverged *in situ* (IIb), and *P. tenuituba* diverged north of the Hengduan Mountains. In Clade 2, seven species/taxa diverged *in situ* (IIb),

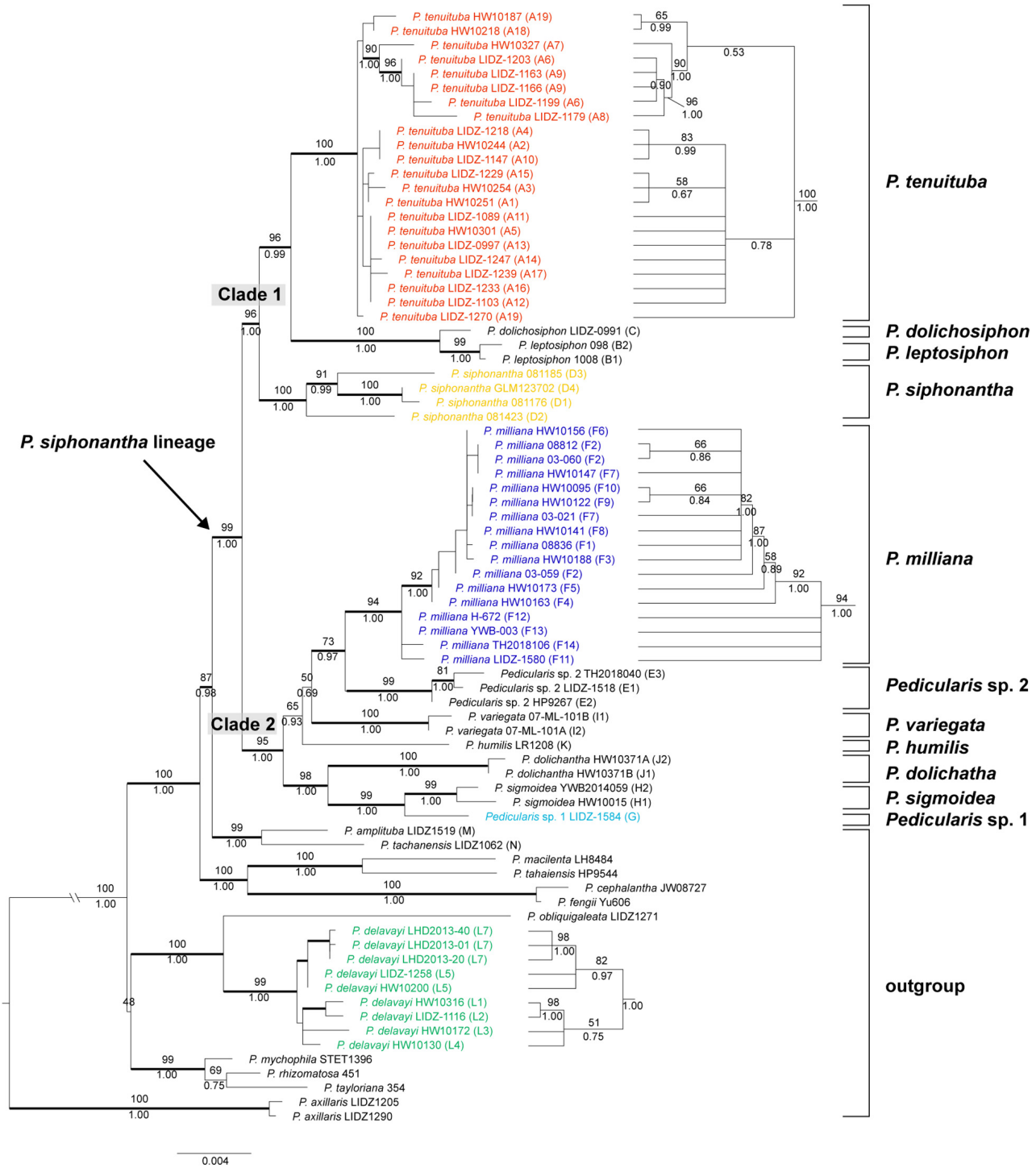


FIGURE 5 | Bayesian inference (BI) tree of the *Pedicularis siphonantha* complex inferred from the modified nrITS+ptDNA dataset by removing the conflicting sequence nrITS of *Pedicularis dolichosiphon* (nrITS) and the four ptDNA regions (*matK*, *rbcL*, *trnH-psbA*, and *trnL-F*) of the samples F11-F14 of *Pedicularis milliana* in accordance with the topological incongruence between the nrITS and ptDNA phylogenies (see **Figure 4** and **Table 3**). Numbers associated with the branches are ML BS value and BI PP, and thicker lines indicate BS ≥ 70 and PP ≥ 0.95. Node with BS < 50 was collapsed. The topology of the *P. milliana* clade with short branch lengths appear on the right.

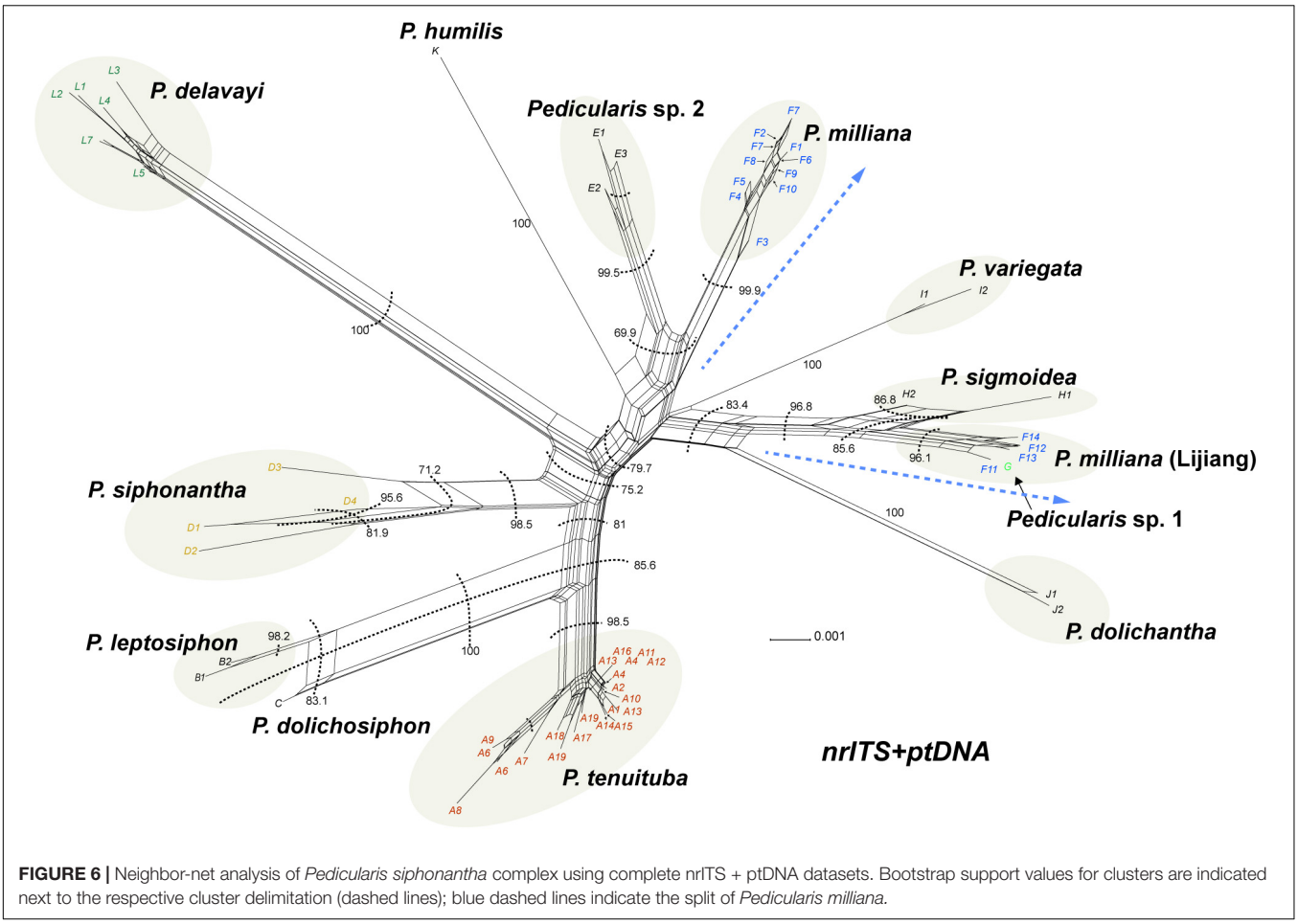


FIGURE 6 | Neighbor-net analysis of *Pedicularis siphonantha* complex using complete nrITS + ptDNA datasets. Bootstrap support values for clusters are indicated next to the respective cluster delimitation (dashed lines); blue dashed lines indicate the split of *Pedicularis milliana*.

with some populations of *P. milliana* migrating northward to IIIb and IIIc.

DISCUSSION

Topological Incongruence Between the nrITS and ptDNA Phylogenies

Topological incongruence between nuclear/nrITS and ptDNA phylogenies has been reported in many taxa (Rieseberg et al., 1996; Buckley et al., 2006; Stegemann et al., 2012; Yi et al., 2015; Stull et al., 2020; Ye et al., 2021). In this study, incongruences between the nrITS and ptDNA datasets were found among species within the *P. siphonantha* complex and in the sister relationship between the *P. siphonantha* complex and *P. amplituba* + *P. tachanensis* in phylogenetic analyses, as well as the estimation of genetic distance. The incongruence could be caused by convergent sequence evolution, incomplete lineage sorting, hybridization/introgression, horizontal gene transfer, and gene duplication/loss (Rokas et al., 2003; Degnan and Rosenberg, 2009). Phylogenetic network analyses suggested that introgression between *P. milliana* and *P. sigmoidea* and between *P. leptosiphon* and *P. tunuituba* is the most plausible explanation for discordance. However, incomplete lineage sorting, convergent sequence evolution, or others cannot be ruled out (Joly et al., 2009).

Introgression might have been common within recently derived species complexes when their distributions overlap (Acosta and Premoli, 2010; Liu et al., 2017; Liu et al., 2020). *Pedicularis* spp. are outcrossed and exclusively pollinated by bumblebees, and pollinator-mediated interspecific gene flow may cause hybridization and introgression among *Pedicularis* species in the same community (Hong and Li, 2005; Yang et al., 2007; Eaton et al., 2012). In the case of *Pedicularis* sect. *Cyathophora*, Yu et al. (2013) have documented that the plastid genome of *P. cyathophylloides* was likely captured from an ancestor of *P. cyathophylla* in the West Sichuan. Similarly, our results show that paraphyletic *P. milliana* populations were associated with distinct geographical ranges, suggesting that either genetic divergence occurred between two clusters due to allopatry and/or a plastid genome capture event. We, therefore, propose an ancient hybridization event between the ancestors of *P. milliana* (♀) and *P. sigmoidea/Pedicularis* sp. 1 (♂) in the Yulong Snow Mountain. In this scenario, high-altitude (≥3,800 m) hybrids (♀) backcrossed with ancestors of *P. milliana* (♀), and low-altitude (<3,800 m) hybrids became established as species at lower altitudes. Therefore, morphological consistency was found among high-altitude populations of *P. milliana*, while low

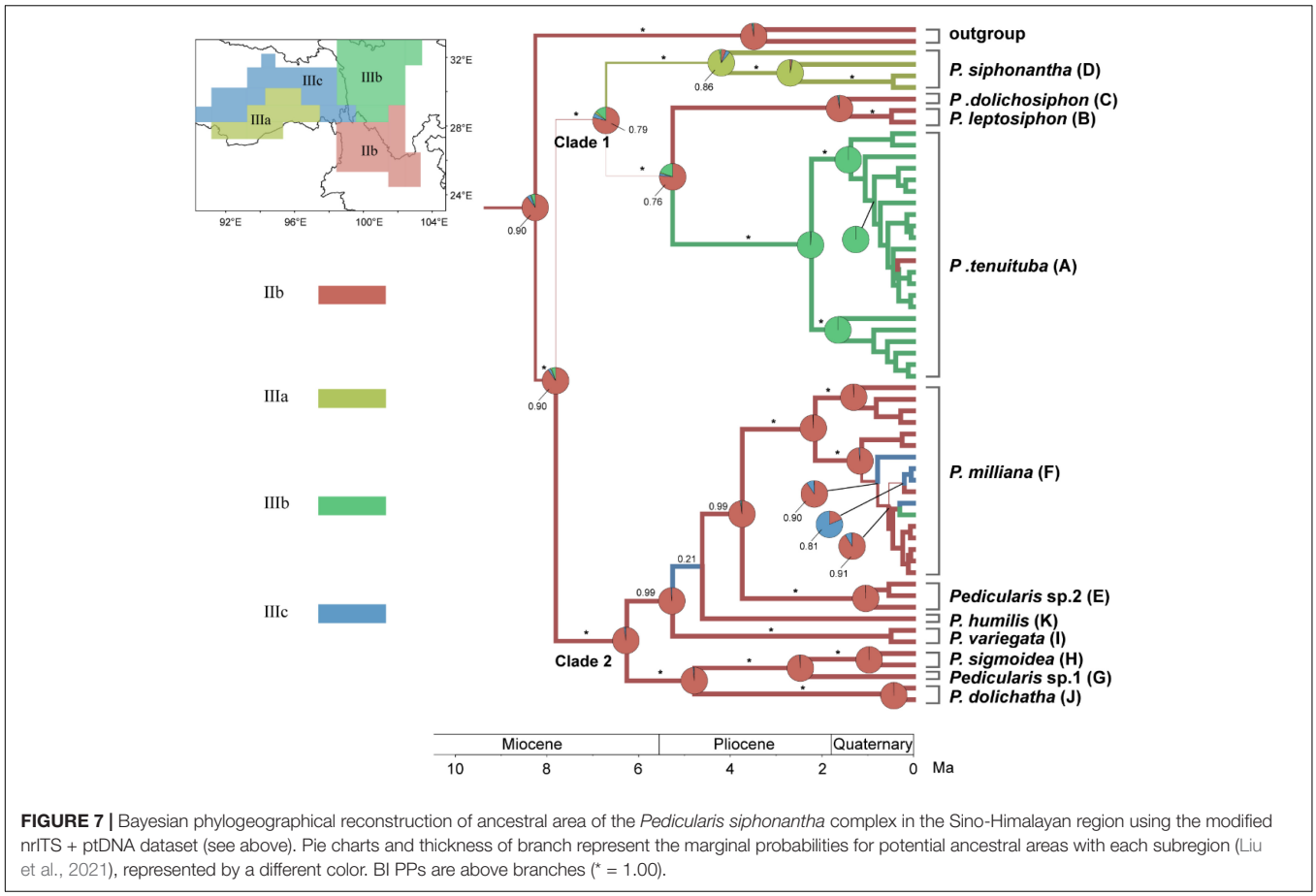


FIGURE 7 | Bayesian phylogeographical reconstruction of ancestral area of the *Pedicularis siphonantha* complex in the Sino-Himalayan region using the modified nrITS + ptDNA dataset (see above). Pie charts and thickness of branch represent the marginal probabilities for potential ancestral areas with each subregion (Liu et al., 2021), represented by a different color. BI PPs are above branches (* = 1.00).

altitude *Pedicularis* sp. 1 diverged from *P. milliana* in the shape of the beak and low lip of corolla. Sample C of *P. dolichosiphon* might also be the result of introgression between *P. tenuituba* (♀) and *P. leptosiphon* (♂), but greater population-level sampling is required to investigate this fully. In addition, more genomic evidence from organelle and nuclear genomes were needed to test these speculations.

Phylogenetic Species Delimitation in the *Pedicularis siphonantha* Complex

Traditionally, the *P. siphonantha* complex together with other long-tubular species belonged to Ser. *Longiflorae* (Li, 1949; Tsoong, 1956). Phylogenetic analyses showed that Ser. *Longiflorae* is polyphyletic, but the *P. siphonantha* complex was monophyletic (Yu et al., 2015, 2018). In this study, ptDNA phylogenies rejected the monophyly of the *P. siphonantha* complex by including two short-tubular species, *P. amplituba* and *P. tachanensis*. The AU and SH tests also could not reject the inclusion of *P. tachanensis* + *P. amplituba* in the *P. siphonantha* complex using the nrITS dataset or the monophyly of the *P. siphonantha* complex using the ptDNA dataset. Moreover, phylogenies of the concatenated nrITS + ptDNA dataset strongly supported the monophyly of the *P. siphonantha* complex (BS/BP = 99/1.00). Therefore, the monophyly of the *P. siphonantha* complex should be accepted, but more nuclear genes and more robust phylogeny should be applied for evaluating this complex in the future.

Although introgression may confound phylogenetic species delimitation of *P. dolichosiphon* and *Pedicularis* sp. 1, species delimitations of the remaining nine species are well-resolved, which is consistent with morphological identification. For *Pedicularis* sp. 2, the corolla beak shape and lower-lip lobes are quite different than its sister species, *P. milliana*. Moreover, *Pedicularis* sp. 2 occurs on the Jiaozi Snow Mountain. Morphological, molecular, and biogeographic evidence all support *Pedicularis* sp. 2 to be a new species. It is worth noting that relatively high intraspecies genetic and phenotypic variations suggest *P. siphonantha* in the Himalayas needs further investigations [also reviewed by Yu et al. (2018)]. The phylogenetic position of *P. humilis* is still not well-resolved; however, it is an isolated species of the Gaoligong Mountain, perhaps the result of allopatric speciation.

Traditionally, morphological character similarity was the main evidence for assessing species relationships, but this criterion might be not suitable in the *P. siphonantha* complex. Because floral characters are labile in *Pedicularis*, morphologically similar species might be only distantly related. For example, phylogenetic analyses showed that *P. milliana* was clustered with morphologically different species including *P. variegata*, *P. humilis*, *P. dolichantha*, and *P. sigmoidea*, rather than the morphologically similar species *P. siphonantha*, contra

previous placements [e.g., Li (1949); Tsoong (1963), and Yang et al. (1998)]. Moreover, the long-tubular species *P. delavayi* (excluded from the *P. siphonantha* complex) was clustered with short-tubular species *Pedicularis obiquigaleata* W. B. Yu and H. Wang. Understanding taxonomic affinities within the *P. siphonantha* complex requires morphological, geographical, and molecular evidence.

Allopatric Speciation in the *Pedicularis siphonantha* Complex

Species diversification of *Pedicularis* in the Mountains of Southwest China is thought to be associated with the uplift of the Qinghai-Tibet Plateau and the establishment of the Asian monsoon climatic cycle (Ding et al., 2020). This rapid diversification resulted in dramatic variation in the form and shape of the corolla (Eaton et al., 2012), though the reasons for several independent transitions of the corolla tube from short to long remain unknown (Huang and Fenster, 2007; Yu et al., 2015, 2018; Huang et al., 2016). Macior and his colleagues (Macior, 1990; Macior and Tang, 1997; Macior et al., 2001) proposed that long-tubular corollas might have some adaptative advantages in the alpine meadow by extending the reproductive organs to attract bumblebee pollinators. The *P. siphonantha* complex, having long-tubular corollas, could be a good model for investigating species diversification in the Mountains of Southwest China. Phylogeographical analysis suggested that ancestors of the *P. siphonantha* complex originated from south of the Hengduan Mountains in the late Miocene, then rapidly expanded westward to the Himalayas, northward of the Hengduan Mountains, and eastward to the Yunnan-Guizhou plateau. Intense orogeny of southern Hengduan Mountains during the late Miocene and Pliocene (Lai et al., 2007; Wang et al., 2012; Favre et al., 2015; Zhang et al., 2017; Farnsworth et al., 2019) likely contributed to environmental heterogeneity driving rapid species divergence in the *P. siphonantha* complex. Therefore, *P. sigmoidea*, *P. dolichantha*, and *Pedicularis* sp. 2 are restricted to the margin of the Hengduan Mountains, while *P. milliana* and *P. tenuituba* in the heartland of the Hengduan Mountains, as well as *P. siphonantha* in the Himalayas, are widely distributed in the vast contiguous alpine meadows of those ranges.

Phylogeny of the modified nrITS + ptDNA dataset resolved the *P. siphonantha* complex as two major clades. In Clade 1, *P. siphonantha* is distributed in the western Himalaya (the type specimen was collected from Nepal), rather than widely distributed in the eastern Himalaya, Sichuan, and Yunnan, as described in the Flora of China (Yang et al., 1998); *P. leptosiphon* and *P. dolichosiphon* are restricted to Ninglang, Yunnan, and Muli, Sichuan; and *P. tenuituba* is widely distributed in western Sichuan, and partly overlaps with *P. leptosiphon* in southwestern Sichuan. Geographic separation may have driven *P. siphonantha* to diverge from the remaining species in this clade. Moreover, the carmine speckles on the lower lobes of *P. tenuituba* may appear distinct from *P. leptosiphon* to bumblebee pollinators where they both occur. It is also worth noting that *P. tenuituba* mainly grows in humid grasslands, but *P. leptosiphon* prefers to grow in the sandy, dry meadows. Niche specification and difference in pollinators may mediate reproductive isolation between *P. tenuituba* and *P. leptosiphon*. However, occasional hybridization might have been responsible for producing the suspected hybrid *P. dolichosiphon*.

The mountainous terrain of the home range of the *P. siphonantha* complex likely maintains genetic isolation among geographically isolated species. For example, *P. dolichantha* and its sister species *P. sigmoidea* are distributed in isolated mountains near Huize and Eryuan, Yunnan, while species closely related to them, *P. humilis* and *P. variegata*, occur in the Gaoligong Mountains and southwest Yunnan and Muli, Sichuan, respectively. *P. milliana* is widely distributed in northwestern Yunnan, and its sister *Pedicularis* sp. 2 (Jiaozi Snow Mountain) is only found in the Jiaozi Snow Mountain, Dongchuan, Yunnan. Of the seven species in this clade, *P. milliana* and *Pedicularis* sp. 1 co-occur in the Yulong Snow Mountain and *P. sigmoidea* has been collected from the south margin of the distribution range of *P. milliana* in Heqing, Eryuan, and Dali. Therefore, geographical isolation likely drove species divergence in this complex, with the exception that introgression between *P. milliana* and *P. sigmoidea* may have produced the suspected hybrid *Pedicularis* sp. 1 and the plastome capture of *P. milliana* from *P. sigmoidea* in Lijiang and south Shangri-La, northwest Yunnan.

CONCLUSION

Overall, phylogenetic analyses of five DNA loci (nrITS, *matK*, *rbcL*, *trnH-psbA*, and *trnL-F*) clarify species delimitation within the *P. siphonantha* complex. Differences in geographical distribution and altitude can be important supplementary indicators to identify species of the *P. siphonantha* complex despite the lack of diagnostic morphological characters in herbarium specimens. The *P. siphonantha* complex likely originated from allopatric speciation. The origin of *P. milliana* and *Pedicularis* sp. 1 in the Lijiang region was plausibly due to an ancestral hybridization event. The morphological, molecular, and biogeographic evidence support taxonomic recognition of *Pedicularis* sp. 2. To better understand the evolution of the *P. siphonantha* complex, further studies of phenotype and environmental factors are needed.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

W-BY, HW, and D-ZL conceived the study. W-BY, HW, J-BY, and RC collected the data. RL and W-BY analyzed the data. RL, W-BY, CR, and D-ZL interpreted the results. All authors wrote and revised the article and approved the final version of the manuscript.

FUNDING

This study was supported by grants from the National Natural Science Foundation of China (31870196 and 32071670), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31000000), and the Ten Thousand Talents Program of Yunnan for Top-notch Young Talents.

ACKNOWLEDGMENTS

We are grateful to Jie Cai, LiNa Dong, Lian-Ming Gao, Hua-Jie He, Wei Jiang, Rong Li, Bin Liu, En-De Liu, Jie Liu, Min-Lu Liu, Lu Lu, Yang Luo, Hui Tang, Chun-Lei Xiang, Ji-Dong Ya, Qiu-Lin Yang, Xiu-Long Yang, and Shu-Dong Zhang for their help in the field work and/or providing plant samples, and to Jing Yang and Zhi-Rong Zhang for their help and suggestions in the lab work, and to the physical support from the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, and the HPC Platform of the Public Technology Service Center, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.811206/full#supplementary-material

Supplementary Figure 1 | Maximum clade credibility tree of *Pedicularis* from BEAST divergence time analysis. The secondary calibration of the Orobanchaceae crown was constrained to 56 ± 10 Mya which was obtained from http://timetree.org/. The estimated age of nodes is presented above the branch. Node bars represent the 95% highest posterior density (HPD) interval.

REFERENCES

Acosta, M. C., and Premoli, A. C. (2010). Evidence of chloroplast capture in South American *Nothofagus* (subgenus *Nothofagus*, Nothofagaceae). *Mol. Phylogenet. Evol.* 54, 235–242. doi: 10.1016/j.ympev.2009.08.008

Alfaro, M. E., Zoller, S., and Lutzoni, F. (2003). Bayes or bootstrap? a simulation study comparing the performance of bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.* 20, 255–266. doi: 10.1093/molbev/msg028

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K., Meier, R., Winker, K., et al. (2007). Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* 22, 148–155. doi: 10.1016/j.tree.2006.11.004

Bouckaert, R. R., and Drummond, A. J. (2017). bModelTest: bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* 17:42. doi: 10.1186/s12862-017-0890-6

Bouckaert, R., Heled, J., Kuehnert, D., Vaughan, T., Wu, C.-H., Xie, D., et al. (2014). BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS Comp. Biol.* 10:e1003537. doi: 10.1371/journal.pcbi.1003537

Boufford, D. E. (2014). Biodiversity hotspot: china's hengduan mountains. *Arnoldia (Jamaica Plain)* 72, 24–35.

Buckley, T. R., Cordeiro, M., Marshall, D. C., and Simon, C. (2006). Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (Maoricicada Dugdale). *Syst. Biol.* 55, 411–425. doi: 10.1080/10635150600697283

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* 37, 291–294. doi: 10.1093/molbev/msz189

Degnan, J. H., and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340. doi: 10.1016/j.tree.2009.01.009

Ding, W.-N., Ree, R. H., Spicer, R. A., and Xing, Y.-W. (2020). Ancient orogenic and monsoon-driven assembly of the world's richest temperate alpine flora. *Science* 369, 578–581. doi: 10.1126/science.abb4484

Don, D., Hamilton, F., and Wallich, N. (1825). *Prodromus Florae Nepalensis: Sive Enumeratio Vegetabilium Quae In Itinere Per Nepaliam Proprie Dictam Et Regiones Conterminas, Ann. 1802-1803*. Londini: J. Gale.

Eaton, D. A. R., Fenster, C. B., Hereford, J., Huang, S.-Q., and Ree, R. H. (2012). Floral diversity and community structure in *Pedicularis* (Orobanchaceae). *Ecology* 93, S182–S194. doi: 10.1890/11-0501.1

Farnsworth, A., Lunt, D. J., Robinson, S. A., Valdes, P. J., Roberts, W. H. G., Clift, P. D., et al. (2019). Past East Asian monsoon evolution controlled by paleogeography, not CO2. *Sci. Adv.* 5:eaax1697. doi: 10.1126/sciadv.aax1697

Favre, A., Michalak, I., Chen, C. H., Wang, J. C., Pringle, J. S., Matuszak, S., et al. (2016). Out-of-Tibet: the spatio-temporal evolution of Gentiana (*Gentianaceae*). *J. Biogeogr.* 43, 1967–1978. doi: 10.1111/jbi.12840

Favre, A., Paeckert, M., Pauls, S. U., Jaehnig, S. C., Uhl, D., Michalak, I., et al. (2015). The role of the uplift of the qinghai-tibetan plateau for the evolution of *Tibetan biotas*. *Biol. Rev.* 90, 236–253. doi: 10.1111/brv.12107

Hebert, P. D. N., and Gregory, T. R. (2005). The promise of DNA barcoding for taxonomy. *Syst. Biol.* 54, 852–859. doi: 10.1080/10635150500354886

Hebert, P. D. N., Cywinska, A., Ball, S. L., and Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proc. Roy. Soc. Lond. B. Biol.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Hillis, D. M., and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42, 182–192. doi: 10.1093/sysbio/42.2.182

Hollingsworth, P. M., Li, D. Z., Van Der Bank, M., and Twyford, A. D. (2016). Telling plant species apart with DNA: from barcodes to genomes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 371:20150338. doi: 10.1098/rstb.2015.0338

Hong, W., and Li, D. Z. (2005). Pollination biology of four *Pedicularis* species (*Scrophulariaceae*) in northwestern Yunnan, China. *Ann. Mo. Bot. Gard.* 92, 127–138.

Hoorn, C., Mosbrugger, V., Mulch, A., and Antonelli, A. (2013). Biodiversity from mountain building. *Nat. Geosci.* 6, 154–154. doi: 10.1038/ngeo1742

Huang, S.-Q., and Fenster, C. B. (2007). Absence of long-proboscid pollinators for long-corolla-tubed Himalayan *Pedicularis* species: implications for the evolution of corolla length. *Int. J. Plant Sci.* 168, 325–331. doi: 10.1086/510209

Huang, S.-Q., Wang, X.-P., and Sun, S.-G. (2016). Are long corolla tubes in Pedicularis driven by pollinator selection? *J. Integr. Plant Biol.* 58, 698–700. doi: 10.1111/jipb.12460

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

IBM (2017). *IBM SPSS Statistics for Windows. Version 25.0*. Armonk, NY: IBM Corp.

Joly, S., Mclenachan, P. A., and Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am. Nat.* 174, E54–E70. doi: 10.1086/600082

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings Bioinf.* 20, 1160–1166. doi: 10.1093/bib/bbx108

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kolaczkowski, B., and Thornton, J. W. (2007). Effects of branch length uncertainty on bayesian posterior probabilities for phylogenetic hypotheses. *Mol. Biol. Evol.* 24, 2108–2118. doi: 10.1093/molbev/msm141

Kress, W. J. (2017). Plant DNA barcodes: applications today and in the future. *J. Syst. Evol.* 55, 291–307. doi: 10.1111/jse.12254

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lai, Q., Ding, L., Wang, H., Yue, Y., and Cai, F. (2007). Constraining the stepwise migration of the eastern Tibetan Plateau margin by apatite fission track thermochronology. *Sci. China Earth Sci.* 50, 172–183. doi: 10.1007/s11430-007-2048-7

Lam-Tung, N., Schmidt, H. A., Von Haeseler, A., and Bui Quang, M. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comp. Biol.* 5:e1000520. doi: 10.1371/journal.pcbi.1000520

Li, H.-L. (1948). A revision of the genus Pedicularis in China. part I. *Proc. Acad. Nat. Sci. Phila.* 100, 205–378.

Li, H.-L. (1949). A revision of the genus Pedicularis in China. part II. *Proc. Acad. Nat. Sci. Phila.* 101, 1–378.

Liu, B.-B., Campbell, C. S., Hong, D.-Y., and Wen, J. (2020). Phylogenetic relationships and chloroplast capture in the Amelanchier-Malacomeles-Peraphyllum clade (Maleae, Rosaceae): Evidence from chloroplast genome and nuclear ribosomal DNA data using genome skimming. *Mol. Phylogenet. Evol.* 147:106784. doi: 10.1016/j.ympev.2020.106784

Liu, M.-L., Yu, W.-B., and Wang, H. (2013). Rapid identification of plant species and iFlora application of DNA barcoding in a large temperate genus *Pedicularis* (Orobanchaceae). *Plant Divers.* 35, 707–714. doi: 10.7677/ynzwyj201313168

Liu, X., Wang, Z., Shao, W., Ye, Z., and Zhang, J. (2017). Phylogenetic and taxonomic status analyses of the Abaso section from multiple nuclear genes and plastid fragments reveal new insights into the North America origin of *Populus* (*Salicaceae*). *Front. Plant Sci.* 7:2022. doi: 10.3389/fpls.2016.02022

Liu, Y., Ye, J.-F., Hu, H.-H., Peng, D.-X., Zhao, L.-N., Lu, L.-M., et al. (2021). Influence of elevation on bioregionalisation: a case study of the Sino-Himalayan flora. *J. Biogeogr.* 48, 2578–2587. doi: 10.1111/jbi.14222

Macior, L. W. (1990). Pollination ecology of Pedicularis punctata Decne. (*Scrophulariaceae*) in the Kashmir Himalaya. *Plant Species Biol.* 5, 215–223. doi: 10.1111/j.1442-1984.1990.tb00181.x

Macior, L. W., and Tang, Y. (1997). A preliminary study of the pollination ecology of *Pedicularis* in the Chinese Himalaya. *Plant Species Biol.* 12, 1–7. doi: 10.1111/j.1442-1984.1997.tb00150.x

Macior, L. W., Tang, Y., and Zhang, J.-C. (2001). Reproductive biology of *Pedicularis* (*Scrophulariaceae*) in the Sichuan Himalaya. *Plant Species Biol.* 16, 83–89. doi: 10.1046/j.1442-1984.2001.00048.x

Maddison, W. P., and Maddison, D. R. (2019). *Mesquite: A Modular System For Evolutionary Analysis. V3.7.0 [Online]*. Available online at: http://www.mesquiteproject.org/ [Accessed August 10, 2020].

Maximowicz, C. J. (1888). Diagnoses plantarum novarum Asiaticarum. *Bull. Acad. Sci. St Petersburg* 32, 427–629.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Ree, R. H. (2005). Phylogeny and the evolution of floral diversity in *Pedicularis* (*Orobanchaceae*). *Int. J. Plant Sci.* 166, 595–613. doi: 10.1086/430191

Rieseberg, L. H., Whitton, J., and Linder, C. R. (1996). Molecular marker incongruence in plant hybrid zones and phylogenetic trees. *Acta Bot. Neerl.* 45, 243–262. doi: 10.1111/j.1438-8677.1996.tb00515.x

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804. doi: 10.1038/nature02053

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Hohna, S., et al. (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508.

Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116. doi: 10.1093/oxfordjournals.molbev.a026201

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

Stegemann, S., Keuthe, M., Greiner, S., and Bock, R. (2012). Horizontal transfer of chloroplast genomes between plant species. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2434–2438. doi: 10.1073/pnas.1114076109

Struck, T. H., Feder, J. L., Bendiksby, M., Birkeland, S., Cerca, J., Gusarov, V. I., et al. (2018). Finding evolutionary processes hidden in cryptic species. *Trends Ecol. Evol.* 33, 153–163. doi: 10.1016/j.tree.2017.11.007

Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A., and Smith, S. A. (2020). Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. *Am. J. Bot.* 107, 790–805. doi: 10.1002/ajb2.1468

Tsoong, P.-C. (1955). A new system for the genus Pedicularis. *Acta Phytotax. Sin.* 4, 71–147.

Tsoong, P.-C. (1956). A new system for the genus Pedicularis. *Acta Phytotax. Sin.* 5, 41–73, 239–278.

Tsoong, P.-C. (1963). "Scrophulariaceae (Pars II)," in *Flora Reipublicae Popularis Sinacae*, Vol. 68, eds S.-S. Chien and W.-Y. Chun (Beijing: Science Press), 61–378.

Wang, E., Kirby, E., Furlong, K. P., Van Soest, M., Xu, G., Shi, X., et al. (2012). Two-phase growth of high topography in eastern Tibet during the Cenozoic. *Nat. Geosci.* 5, 640–645. doi: 10.1038/ngeo1538

Xing, Y., and Ree, R. H. (2017). Uplift-driven diversification in the Hengduan Mountains, a temperate biodiversity hotspot. *Proc. Natl. Acad. Sci. U.S.A.* 114, E3444–E3451. doi: 10.1073/pnas.1616063114

Yang, C.-F., Gituru, R. W., and Guo, Y.-H. (2007). Reproductive isolation of two sympatric louseworts, *Pedicularis rhinanthoides* and *Pedicularis longiflora* (Orobanchaceae): how does the same pollinator type avoid interspecific pollen transfer? *Biol. J. Linn. Soc.* 90, 37–48. doi: 10.1111/j.1095-8312.2007.00709.x

Yang, F. S., Qin, A. L., Li, Y. F., and Wang, X. Q. (2012). Great genetic differentiation among populations of *Meconopsis integrifolia* and its implication for plant speciation in the Qinghai-Tibetan Plateau. *PLoS One* 7:e37196. doi: 10.1371/journal.pone.0037196

Yang, H.-B., Holmgren, N. H., and Mill, R. R. (1998). "Pedicularis Linn," in *Flora of China*, Vol. 18, eds Z.-Y. Wu and P. H. Raven (St. Louis, MI: Missouri Botanical Garden Press & Science Press), 97–209. doi: 10.1016/j.biopha.2017.10.133

Ye, X. Y., Ma, P. F., Guo, C., and Li, D. Z. (2021). Phylogenomics of *Fargesia* and *Yushania* reveals a history of reticulate evolution. *J. Syst. Evol.* 59, 1183–1197. doi: 10.1111/jse.12719

Yi, T.-S., Jin, G.-H., and Wen, J. (2015). Chloroplast capture and intra- and inter-continental biogeographic diversification in the Asian – New World disjunct plant genus Osmorhiza (*Apiaceae*). *Mol. Phylogenet. Evol.* 85, 10–21. doi: 10.1016/j.ympev.2014.09.028

Yu, W. B., Huang, P. H., Li, D. Z., and Wang, H. (2013). Incongruence between nuclear and chloroplast DNA phylogenies in *Pedicularis* section *Cyathophora* (*Orobanchaceae*). *PLoS One* 8:e74828. doi: 10.1371/journal.pone.0074828

Yu, W. B., Huang, P.-H., Ree, R. H., Liu, M.-L., Li, D.-Z., and Wang, H. (2011). DNA barcoding of *Pedicularis* L. (*Orobanchaceae*): evaluating four universal barcode loci in a large and hemiparasitic genus. *J. Syst. Evol.* 49, 425–437. doi: 10.1111/j.1759-6831.2011.00154.x

Yu, W. B., Liu, M. L., Wang, H., Mill, R. R., Ree, R. H., Yang, J. B., et al. (2015). Towards a comprehensive phylogeny of the large temperate genus *Pedicularis* (*Orobanchaceae*), with an emphasis on species from the Himalaya-Hengduan Mountains. *BMC Plant Biol.* 15:176. doi: 10.1186/s12870-015-0547-9

Yu, W. B., Wang, H., Liu, M. L., Grabovskaya-Borodina, A. E., and Li, D. Z. (2018). Phylogenetic approaches resolve taxonomical confusion in *Pedicularis* (*Orobanchaceae*): reinstatement of *Pedicularis delavayi* and discovering a new species *Pedicularis milliana*. *PLoS One* 13:e0200372. doi: 10.1371/journal.pone.0200372

Zha, H.-G., Milne, R. I., and Sun, H. (2010). Asymmetric hybridization in Rhododendron agastum: a hybrid taxon comprising mainly F(1)s in Yunnan, China. *Ann. Bot.* 105, 89–100. doi: 10.1093/aob/mcp267

Zhang, D.-C., Ye, J.-X., and Sun, H. (2016). Quantitative approaches to identify floristic units and centres of species endemism in the Qinghai-Tibetan Plateau, south-western China. *J. Biogeogr.* 43, 2465–2476. doi: 10.1111/jbi.12819

Zhang, Y.-Z., Replumaz, A., Leloup, P. H., Wang, G.-C., Bernet, M., Van Der Beek, P., et al. (2017). Cooling history of the Gongga batholith: implications for the Xianshuihe Fault and Miocene kinematics of SE Tibet. *Earth Planet. Sci. Lett.* 465, 1–15. doi: 10.1016/j.epsl.2017.02.025

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer C-LX declared a shared affiliation, with no collaboration, with several of the authors HW, J-BY, and D-ZL to the handling editor at the time of the review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 11 April 2022
doi: 10.3389/fpls.2022.808156

Comparative Analyses of 3,654 Plastid Genomes Unravel Insights Into Evolutionary Dynamics and Phylogenetic Discordance of Green Plants

Ting Yang[1,2,3†], Sunil Kumar Sahu[1,2†], Lingxiao Yang[4], Yang Liu[1,2], Weixue Mu[1,2], Xin Liu[1,2], Mikael Lenz Strube[3], Huan Liu[1,2,5] and Bojian Zhong[4*]*

[1] Beijing Genomics Institute Shenzhen, Yantian Beishan Industrial Zone, Shenzhen, China, [2] State Key Laboratory of Agricultural Genomics, Beijing Genomics Institute Shenzhen, Shenzhen, China, [3] Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark, [4] College of Life Sciences, Nanjing Normal University, Nanjing, China, [5] Department of Biology, University of Copenhagen, Copenhagen, Denmark

The plastid organelle is essential for many vital cellular processes and the growth and development of plants. The availability of a large number of complete plastid genomes could be effectively utilized to understand the evolution of the plastid genomes and phylogenetic relationships among plants. We comprehensively analyzed the plastid genomes of Viridiplantae comprising 3,654 taxa from 298 families and 111 orders and compared the genomic organizations in their plastid genomic DNA among major clades, which include gene gain/loss, gene copy number, GC content, and gene blocks. We discovered that some important genes that exhibit similar functions likely formed gene blocks, such as the *psb* family presumably showing co-occurrence and forming gene blocks in Viridiplantae. The inverted repeats (IRs) in plastid genomes have doubled in size across land plants, and their GC content is substantially higher than non-IR genes. By employing three different data sets [all nucleotide positions (nt123), only the first and second codon positions (nt12), and amino acids (AA)], our phylogenomic analyses revealed Chlorokybales + Mesostigmatales as the earliest-branching lineage of streptophytes. Hornworts, mosses, and liverworts forming a monophylum were identified as the sister lineage of tracheophytes. Based on nt12 and AA data sets, monocots, Chloranthales and magnoliids are successive sister lineages to the eudicots + Ceratophyllales clade. The comprehensive taxon sampling and analysis of different data sets from plastid genomes recovered well-supported relationships of green plants, thereby contributing to resolving some long-standing uncertainties in the plant phylogeny.

Keywords: plastid genome, phylogenetics, Viridiplantae, inverted repeats, gene blocks

OPEN ACCESS

Edited by:
Stefan Wanke, Technical University Dresden, Germany

Reviewed by:
Juan Carlos Villarreal A., Laval University, Canada
Shiou Yih Lee, INTI International University, Malaysia
Wei Lun Ng, Xiamen University Malaysia, Malaysia

***Correspondence:**
Sunil Kumar Sahu
sunilkumarsahu@genomics.cn
Bojian Zhong
bjzhong@gmail.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *03 November 2021*
Accepted: *07 March 2022*
Published: *11 April 2022*

Citation:
Yang T, Sahu SK, Yang L, Liu Y, Mu W, Liu X, Strube ML, Liu H and Zhong B (2022) Comparative Analyses of 3,654 Plastid Genomes Unravel Insights Into Evolutionary Dynamics and Phylogenetic Discordance of Green Plants. Front. Plant Sci. 13:808156. doi: 10.3389/fpls.2022.808156

INTRODUCTION

Chloroplasts are the defining organelle of the plant lineage, essential for photosynthesis, lipid metabolism, and innumerable other cellular processes related to plant growth, development, and stress response. Since the endosymbiotic origin of plastids, gene transfer from the plastid genome (plastome) to the nucleus is a continuous process (Matsuo et al., 2005; Eckardt, 2006). Therefore, phylogenetic trees based on a few plastid genes may lead to incongruence. However, plastid genomic DNA (ptDNA) is conserved in gene content (Wicke et al., 2011). The conserved plastid gene blocks could be explained by large-scale gene transfers in an ancestral lineage, among others. For instance, the presence of gene blocks such as *psbB/T/N/H* could be considered as an indication of monophyly of streptophytes (Lee and Manhart, 2002; Howe et al., 2008).

Plastid DNA of green plants (Viridiplantae) normally exhibits a conserved genome structure, which contains two copies of an inverted repeat (IR) separating a small single-copy (SSC) region from the large single-copy region (LSC). The plastome sizes of photosynthetic land plants normally range from 107 (*Cathaya argyrophylla*, Pinaceae) (Lin et al., 2010) to 218 kb (*Pelargonium*, Geraniaceae) (Chumley et al., 2006). However, some angiosperm lineages may have extreme variations in their genome size (Wicke and Naumann, 2018; Chen et al., 2020; Lyko and Wicke, 2021; Li et al., 2022). For instance, the plastid genomes of parasitic plants such as *Pilostyles* spp. or *Prosopanche americana* (Hydnoraceae) are only around 12 and 28 kb, respectively (Bellot et al., 2016; Arias-Agudelo et al., 2019; Jost et al., 2020). In contrast, the plastid genomes of the chlorophyte *Floydiella* (Chaetopeltidaceae) is 520 kb in length (Brouard et al., 2010). The sizes of plastic genomes (ptDNA) have been compared within many clades (Xu et al., 2015; Xiao-Ming et al., 2017). Many factors are known to cause plastome size variation, which includes (a) variations of intergenic regions, and intron lengths (Maul et al., 2002; Simpson and Stern, 2002), (b) IR region variation (Chumley et al., 2006; Brázda et al., 2018), and (c) gene loss (Braukmann et al., 2013; Chen et al., 2020; Jost et al., 2020). An IR analysis of all green plants showed that shorter IRs are frequently found in bryophytes followed by chlorophytes, while Polypodiopsida with the lowest frequencies (Brázda et al., 2018). However, in Papilionoideae, Pinaceae, and cupressophytes, the IRs are nearly lost or missing (Wu et al., 2011; Lin et al., 2012; Xu et al., 2015), with at least two independent regains of IRs following a previous loss (Choi et al., 2019; Qu et al., 2019). Gene content variation contributes to the plastome size variation only to a smaller extent, with an exception of heterotrophic algae and parasitic flowering plants, which have partially or completely lost their photosynthetic ability (Wicke and Naumann, 2018; Lyko and Wicke, 2021).

To understand the origin and relationships of green plants, the phylogenetic analyses have been widely performed based on nuclear (e.g., Wickett et al., 2014; One Thousand Plant Transcriptomes Initiative, 2019), mitochondrial (Liu et al., 2014), and plastid loci (Nickrent et al., 2000; Burleigh and Mathews, 2004; Li et al., 2019, 2022; Sousa et al., 2020). The phylogenetic relationship among chlorophytes has been reviewed recently (Leliaert et al., 2011, 2012; Lemieux et al., 2016; Fang et al., 2017; Li et al., 2020). However, the relationships among core chlorophyte clades (Chlorodendrophyceae, Ulvophyceae, Trebouxiophyceae, and Chlorophyceae) require further analyses (Li et al., 2021b). Large-scale transcriptome data resolved topological uncertainty within ferns and bryophytes (Pryer et al., 2004; Shaw and Renzaglia, 2004; Shen et al., 2017; Puttick et al., 2018; Sousa et al., 2020). Lu et al. (2014) used two nuclear genes and performed near-complete sampling of extant gymnosperms genera and found that cycads are the basal-most lineage of gymnosperms rather than a sister group to Ginkgoaceae (Lu et al., 2014). Burleigh and Mathews (2004) used four nuclear loci, five chloroplast loci, and four mitochondrial loci from 31 genera to resolve the seed plant tree of life (Burleigh and Mathews, 2004). Another group used 61 plastid genes from 45 taxa to reconstruct the phylogenetic order among basal angiosperms (Moore et al., 2007). A nearly complete set of plastid protein-coding sequences based on 360 species of the green plants (Gitzendanner et al., 2018) and 1,879 taxa representing all the major subclades across green plant have been reported (Ruhfel et al., 2014). Likewise, the large-scale phylogenomic study using 1,342 transcriptomes that represent 1,124 species has been performed across green plants (One Thousand Plant Transcriptomes Initiative, 2019). Despite the expanded taxon sampling and comprehensive plastome data set, relationships among the five major clades of Mesangiospermae remain elusive (Li et al., 2021a).

Next-generation sequencing technologies have contributed to complete plastid genomes of plants. Until January 2021, over 3,823 complete plastid genome sequences have been published in the National Center for Biotechnology Information (NCBI) organelle genome database. This large amount of complete ptDNA data can be effectively utilized to understand the evolution of plastid genomes and infer phylogenetic relationships among plants. By employing these large-scale data, we aimed to understand (i) the overview of the plastome architecture in Viridiplantae following the split from chlorophytes, and phylogenetic relationships mainly focusing on core chlorophytes, ferns and bryophytes, Mesangiospermae (comprising magnoliids, Chloranthales, monocots, Ceratophyllum, and eudicots) based on nt12, nt123, AA of plastid protein-coding genes, (ii) how the gene order (positional arrangement) is shaped along the Viridiplantae, (iii) what forces could underly the formation and uneven size distribution of IRs in Viridiplantae, and (iv) whether an increased taxon sampling helps to resolve phylogenetic relationships and topological conflicts in Viridiplantae. To answer these questions, we analyzed plastid genome data from 3,654 taxa, 298 families, 111 orders of Viridiplantae and compared the genomic organizations in their ptDNAs, which include gene gains/losses, gene copy number variation, GC content, and plastid gene blocks. We also covered a wide range of green plant species to infer plastid data-based phylogenetic trees and compared to previously phylogenomic analyses. The analyses based on wide coverage in taxon sampling allowed us to gain new insights into evolutionary dynamics and the phylogeny of Viridiplantae.

RESULTS AND DISCUSSION

The Genome Size and Gene Organization in Plastid Genomes

In this study, the complete plastid genomes (ptDNA) of 3,654 taxa (available as of Jan 2019), which represent 298 families, and 111 orders of Viridiplantae were selected, comprising chlorophytes (70), charophytes (12), liverworts (6), mosses (8), hornworts (2), lycophytes (5), ferns (85), gymnosperms (202), and angiosperms (3,264) (**Supplementary Table 1**). The size of ptDNA ranged from 521,168 to 71,666 bp. Liverworts, mosses, and gymnosperms displayed the smallest average genome size, which was 118.26, 129.08, and 127.53 kb, respectively, whereas chlorophytes had the largest genome size variation with an average genome size of 156.23 kb (**Figure 1**).

Even though plastid genome sizes show large variation, gene numbers are rather conserved comprising 120–130 genes. We recovered 72 protein-coding genes from all the sequenced ptDNA (seven genes: *ndhF*, *psaA*, *psaB*, *rpoB*, *rpoC1*, *rpoC2*, and *ycf2* were not included in this study, refer to section "Materials and Methods"), and to investigate the status of gene content in the Viridiplantae, we calculated the average gene number in every order to investigate the status of gene content in the Viridiplantae. The overview of the genes is presented in **Supplementary Figure 1**. We found that most of the protein-coding genes normally present as a single copy. Most of the chlorophytes, the gymnosperm order Gnetales and Pinales, and the eudicot Santalales harbor no genes corresponding to the *ndh* family. All angiosperms have *ndh* genes and possess two copies of *rps12, rpl2*, *rps7*, and *rpl23*, as well as *ndhB*. Similarly, the number of introns in ptDNA of Viridiplantae is generally conserved (**Figure 1**). Most of the genes lacked introns with the exception among several ribosomal proteins and photosynthesis genes (**Supplementary Table 1**). The genes that include *atpF*, *ndhA*, *ndhB*, *petB*, *petD*, *rpl16*, *rps12*, *rps16,* and *ycf3* possessed one intron in most of Streptophyta. The intron number of *clpP* gene showed a high divergence, with 2,327 species having two introns and more than 100 species having 3–4 introns. But no intron was found in *clpP* among chlorophytes, gymnosperms (except Ginkgoales and Cycadales), and Poaceae of monocots.

The GC bias is widely discovered in the plastid genomes (e.g., Ruhfel et al., 2014; Chen et al., 2021). In this study, we constructed different data sets to calculate the GC content between 14 major clades of Viridiplantae (**Figure 2**). Specifically, we used five sets of 72 protein-coding genes, from which we carried over the first base (GC1), the second base (GC2), and the third base (GC3) of codon and subjected them with a ntNo3rd (GC12) for the GC content analysis. The total GC content ranged from 34.0 to 42.2% in chlorophytes, 36.8–39.7% in charophytes, 36.7–42.5% in bryophytes, 41.1–56.8% in lycophytes, 37.5–45.4% in fern, around 41.0% in gymnosperms, and 39.9–41.4% in angiosperms (**Figure 2A**). There was a non-significant difference in GC content among seed plant, despite the fact that lycophytes had a significantly greater GC content. Many plastid genomes have revealed that the GC content at each base of the codon is different and GC1 > GC2 > GC3 (e.g., Kim et al., 2014; Zhang et al., 2016). According to the results of the GC content analyses, the GC3 had significantly lower values for all 14 clades, with particularly low values for charophytes, chlorophytes, and bryophytes (**Figure 2B**). The previous analyses have shown that genes in the conserved order tend to evolve more slowly and with a higher proportion of GC than genes in the non-conserved order in bacteria (Papanikolaou et al., 2009). The *psb* family are important plastid genes which encode photosystem II proteins. In our study, we found that *psbB-psbT-psbN-psbH* always appeared in one cluster, and each gene had a consistent GC content throughout the 14 clades (**Supplementary Figure 2**). The average GC content for the *psbB-psbT-psbN-psbH* gene family was 42.04%, whereas the average GC content for the non-conserved *psb* family (*psbA*, *psbI*, *psbK*, and *psbL*) was only 33.81%. Not only the order of gene conservation can affect the GC content, but also the selection and recombination shaped it. For instance, GC content is known to increase rapidly in recombination hotspots (Meunier and Duret, 2004; Marsolier-Kergoat and Yeramian, 2009; Sundararajan et al., 2016). The previous studies have also shown that genes relocated to IRs tend to gain high GC content (Wu and Chaw, 2015; Li et al., 2016). Therefore, we compared the GC content changes in five genes (*rps19*, *rps2*, *rpl23*, *rps7*, and *ndhB*), which underwent twofold expansion in the IRs. A number of five genes were classified as "in-IRs" when found in IR regions, whereas the others were classified as "out-IRs" when they are absent in IR regions. With the exception of *rps19*, we observed a significant variation in GC content and also made an interesting observation that genes that were transported into IRs are likely to have higher GC content than genes that were not transported into IRs (**Figure 2D**).

Gene Loss/Gain in Plastid Genomes and Dynamic Evolution of Inverted Repeat in Green Plants

Although the genetic content and number of protein-coding genes are generally conserved in the plastid genomes, gene gains and losses have been reported in the previous analyses (Gao et al., 2010; Wicke et al., 2011; Mohanta et al., 2020).

The functional role of *ndh* genes is intimately connected with the adaptability of terrestrial plants and photosynthesis (Papanikolaou et al., 2009; Martín and Sabater, 2010). In this study, *ndh* genes were found to be lost in at least 300 species. The *ndh* genes are absent in all plastid DNAs of chlorophytes except Palmophyllales and Pyramimonadales. With the exception of Pinaceae, Gnetales, Erodium, and most Orchidaceae, the plastid DNAs of Streptophyta contain the *ndh* genes. However, in Campanulaceae, Ericaceae, and Fabaceae, *ndh* genes were found to be duplicated. At the same time, except *ndh* gene family, *petN*, *matK*, *rpl22*, *rpl33*, *rps15*, and *rps16* were lost in chlorophytes. We found that some genes are more likely to be lost in some streptophytes. For example, *infA* was absent in 1,825 taxa, and it was more frequently observed among angiosperms, especially in eudicots; *ycf1* and *accD* were missing in more than 800 taxa in angiosperms, especially in monocots; *rpl22*, *rps16*, *ycf1*, *ycf4*, and *infA* are widely absent in Fabaceae (**Supplementary Table 1**). Genes lost from the plastid genome may have moved to the

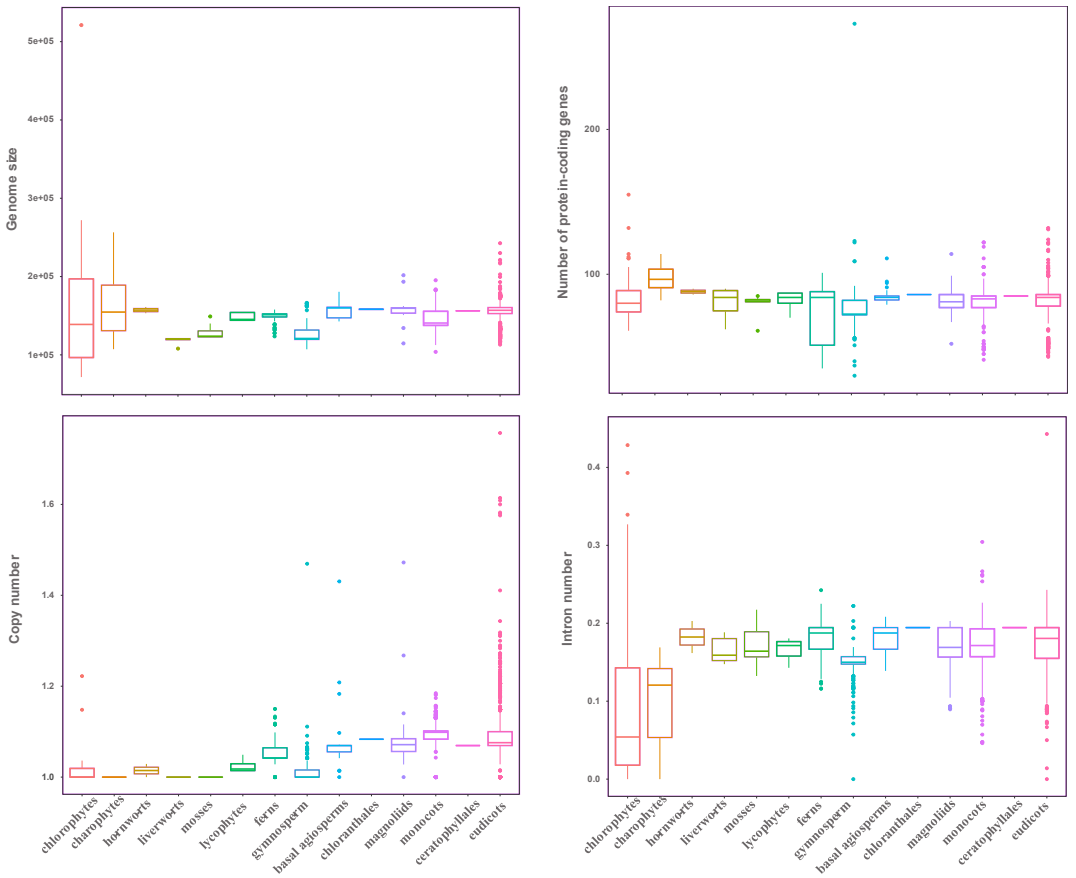


FIGURE 1 | Characteristic features of plastid genomes. The genome size, protein-coding genes number, gene copy number, and intron number in Viridiplantae. Boxplots represent minimum, median, and maximum values.

nuclear or been replaced by related proteins, such as *infA* (Millen et al., 2001), *rpl22,* and *rps16* (Keller et al., 2017), but some are predicted to be indispensable under favorable conditions, such as *ndh* genes (Ruhlman et al., 2015).

The plastid genomes display a quadripartite structure and carry two identical copies of a large IR in all green plants. Some researchers believed that a pair of large IR could stabilize the plastid genome against major structural rearrangements (Strauss et al., 1988; Wu and Chaw, 2014). IRs in green algae showed large fluctuation in size from 6.8 to 45.5 kb and sustained losses in major groups of green algal. For example, *Ulva* (Liu and Melton, 2021), Bryopsidales (Cremen et al., 2018), and Chlorellales (Turmel et al., 2009) lack the IR regions. Some members of Ulvophyceae and Ulvales do have IRs which encode the rRNA, but gene contents and gene orders showed greater diversity. Even though the quadripartite structure shows a high degree of conservation in land plants, but the boundaries of IRs changed significantly in the land plants. The acquisitions of genes by IR expansions have repeatedly been documented (e.g., Wang et al., 2008; Zhu et al., 2016). During land plant evolution, the expansion of IRs from the SC regions has occurred at least two times (Waltari and Edwards, 2002). IRs normally contain tRNAs and rRNAs, but we did not annotate tRNA and rRNAs; instead, we mainly focused on six coding genes (*rps19, rpl2, rpl23, ndhB, rps7, and rps12*) which were widely present in the IRs of angiosperms (**Supplementary Table 2**). Across land plants, the terminal IR gene (IRA) adjacent to the LSC region was observed to be highly conserved (*psbB-psbT-psbN-psbH-petB-petD-rpoA-rps11-rpl36-infA-rps8-rpl14-rpl16-rps3-rpl22*) (**Supplementary Figure 3**). *ndhB-rps7-rps12* and *rps19-rpl2-rpl23-ndhB-rps7-rps12* were newly acquired in IRs of seed plants and angiosperms, respectively. The *rps19-rpl2-rpl23* were conserved in the green plants, but *ndhB-rps7-rps12* showed greater variation. With some duplications, *ndhB/rps7/rps12* in some hornworts exist at the end of LSC and are connected with IRB. In lycophytes, the IR region showed a minor expansion, where *ndhB, rps7*, and *rps12* were expanded to IRs (the first-time expansion). Notably, for the first time, the exon 2 of *rps12*; *rps7*, *ndhB*; *rps7*, and exons 2–3 of *rps12* and *ndhF* were added to the IRs of *Huperzia, Isoetes*, and *Selaginella*, respectively (Wolf et al., 2005; Mower et al., 2019). Based on the structural evolution of Lycopodiaceae plastome and the position of *ndhB, rps7, and rps12*, we hypothesized that the IR expansion was associated with structural inversion and duplication of *ndhB, rps7, and rps12* near IRB, followed by the inversion into junction between the highly conserved IRA region. In ferns, except *rps19-rpl2-rpl23-ndhB-rps7-rps12* block

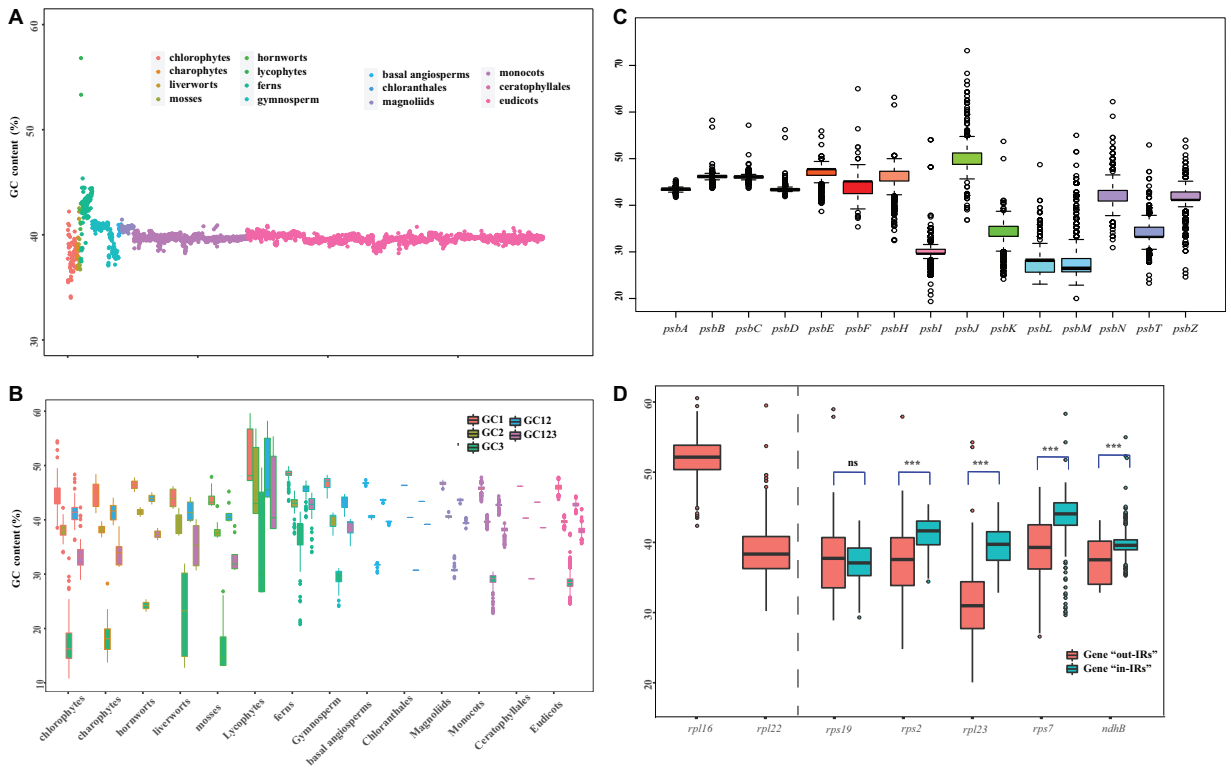


FIGURE 2 | Overview of GC content in Viridiplantae. **(A)** GC content variation among the 14 major lineages of Viridiplantae. **(B)** GC content variation based on five sets of 72 protein-coding genes represented by first base (GC1), the second base (GC2), the third base (GC3) of codon, along with GC123 and GC12. **(C)** GC content variation in *psb* family genes. **(D)** GC content variation of five genes located in IR and non-IR region. Boxplots represent minimum, median, and maximum of GC content. Asterisks (*) represent the significant difference from respective genes using Student's *t*-test (***$p < 0.001$); *ns* = not significant.

in Marattiales, most orders have *ndhB-rps12-rps7-psbA-ycf1* block, which is near the IR regions. In angiosperms, almost all the flowering plants exhibited IR expansion and gained two copies of *rps19, rpl2, rpl23, ndhB, rps7, and rps12* (the second-time expansion), especially in Nymphaeales, about nine to 20 genes in LSC expanded into the IRA compared to Amborellales and then were duplicated in the IRB region.

Gene Conservation and Gene Blocks

It is well known that the structure of plastid genomes is conserved and the order (positional arrangement) of genes is relatively consistent in land plants. This opens up the possibility of reconstructing insertions, deletions, and inversions during the evolution of green plants. In this study, 72 protein-coding genes were ordered according to the annotated position. In *Arabidopsis thaliana*, block analysis has been done based on chloroplast transcriptome expression, and the chloroplast genes are grouped into eight subblocks (Geimer et al., 2008). To calculate the blocks' frequency in Streptophyta, we first removed the samples that showed similar gene content at the order level and finally obtained 1,517 ptDNA. The blocks' frequencies are listed in **Supplementary Table 3**. We found that the classes exhibiting similar functions likely formed gene blocks, with ATP synthase, Phytosystem, and Cytochrome as well as Ribosomal block appearing more than one time with high frequency. Based on the functional categories, there were three major gene blocks. The frequency of ATP synthase block: *atpA-atpF-atpH-atpI* was 74% and *atpE-atpB* was 82%; in Phytosystem and Cytochrome: *petA-psbJ-psbL-psbF-psbE-petL-petG* was 80%, *psbB-psbT-psbN-psbH-petB-petD* was 85%; and in Ribosomal: *rps8-rpl14-rpl16-rps3* was 83%, *rpl33-rps18-rpl20* was 82%, and *rpoA_rps11_rpl36* was 85%. In monocots and eudicots, we observed three photosystem gene blocks with high frequency: *psbM/D/C/Z* [60%], *psbJ/L/F/E* [85%], and *psbB/T/N/H* [88%]. *PsbJ/L/F/E* and *psbB/T/N/H* were nearly conserved in all the green plants and putatively formed blocks: *psbB/T/N/H-petB-petD-rpoA-rps11-rpl36* [78%], *psbJ/L/F/E-petL-petG-psaJ-rpl33-rps18-rpl20* [76%] in Streptophyta. Interestingly, in *A. thaliana*, *psbB/T/N/H-petB-petD* and *rps3-rpl22-rps19-rps2-rps23* show similar gene expression pattern, which is quite different from *rpoA-rps11-rpl36-rps8-rpl14-rpl16* under various biological conditions (Geimer et al., 2008). However, *psbM/D/C/Z* block showed the highest variability in Viridiplantae. *PsbD* and *psbC* genes encode the D2 and CP43 proteins of the photosystem II complex, and they are generally co-transcribed (Adachi et al., 2011). Similarly, *psbM* is highly light-sensitive and plays an important role in such conditions; in fact, the knockout of *psbM* leads to a significant decrease in the activity of photosystem II (Umate et al., 2007). In chlorophytes, *psbD/C/Z*, *psbZ/M*, and *psbD/C* were found to be widely distributed, but in charophytes,

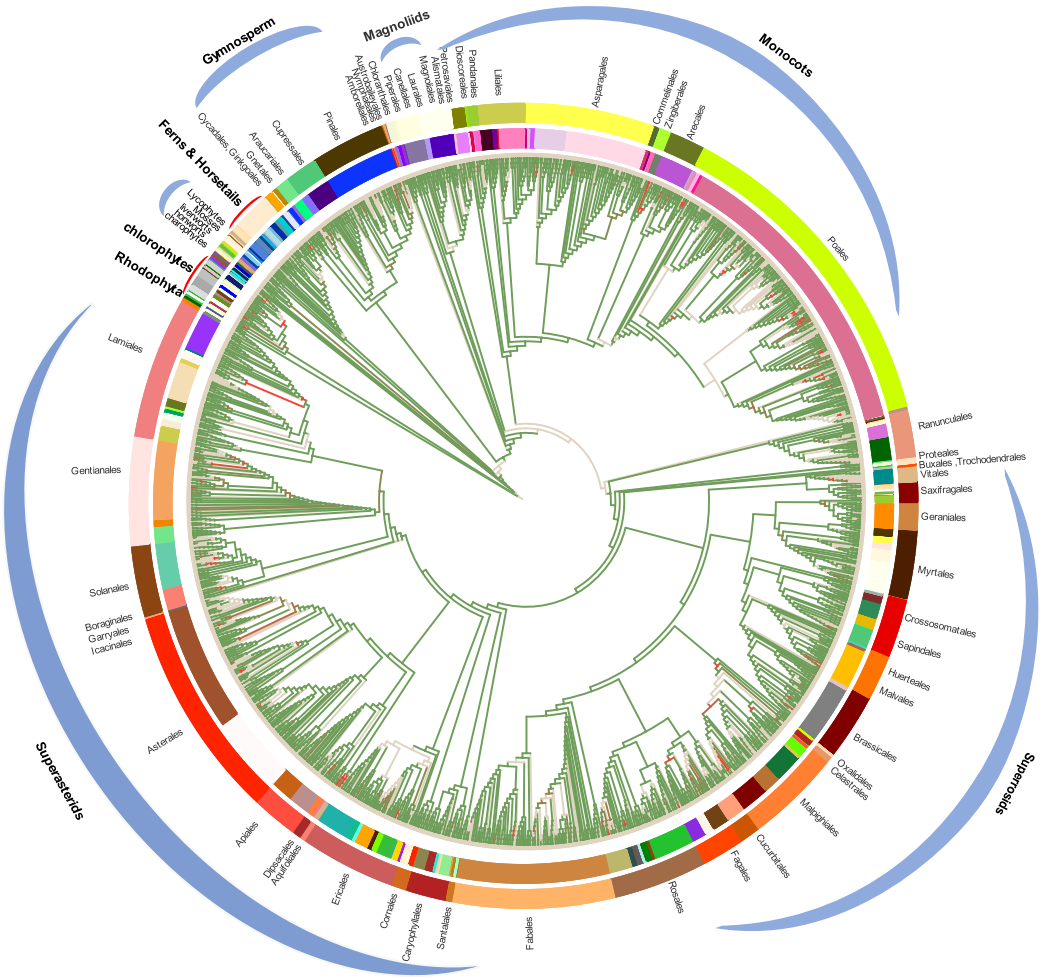


FIGURE 3 | Plastid phylogenomic tree inferred based on the matrix nt12 of 72 protein-coding genes of 3,654 green plants and six Rhodophyta using IQTREE. The colors in the internal circle indicate different families whereas the colors in the external circle indicate different orders (Further details can be found in **Supplementary Figure 11**). The green branches represent the branch with more than 95% UFboot.

only *psbD/C/Z* block exists. Later in bryophytes, *psbZ/C/D* and *psbM* were connected by ATP synthase: *atpA/F/H/I*. For ferns and horsetails clade, the block of *psbM/D/C/Z* was formed. In Cycadales, complete *psbM/D/C/Z* blocks were retained, but *psbM* and *psbD/C/Z* were separated in Pinales. In Poaceae, *atpA/F/H/I-rps2-petN-psbM* was especially inverted, which leads to the production of larger block *psbK/I/M/D/C/Z*.

Except for gene blocks for specific classes that exhibit similar functions, there were several large blocks having more than one functional category genes that exhibit different frequencies. The largest block: (*atpA-atpF-atpH-atpI*) - (*rps2-petN-psbM*) - (*psbD-psbC-psbZ*) - (*rps14-ycf3-rps4*) [51%]- (*ndbJ-ndhK-ndhC-atpE-atpB-rbcL*) [70%] *-accD-psaI-*(*ycf4-cemA-petA-psbJ-psbL-psbF-psbE-petL-petG-psaJ-rpl33-rps18-rpl20*) [69%]- (*psbB-psbT-psbN-psbH-petB-petD-rpoA-rps11-rpl36*) [78%] was found with high frequency in Streptophyta (numbers in [] are the block frequency). In Streptophyta, the block: (*psbB-psbT-psbN-psbH-petB-petD*) [85%] -(*rpoA-rps11-rpl36*) [85%]-*infA*-(*rps8-rpl14-rpl16-rps3-rpl22-rps19-rps2-rps23*) [61%] widely existed and was located near IR regions. Parts of this block are the S10–spc–alpha operon locus that first appeared in eubacteria (Coenye and Vandamme, 2005). The S10-spc regions in the *Euglena* and glaucophyte plastids contained *rpl23-rpl2-rps19-rpl22-rps3-rpl16-rps17-rpl14-rpl5-rps8* (Figueroa-Martinez et al., 2019), which were identical to that in the *E. coli* operons (Clark, 2013). Even in prokaryotic genomes (Coenye and Vandamme, 2005), this location in ptDNA might be derived from these prokaryotes to Viridiplantae.

Congruence and Conflict in Phylogenetic Trees

To conduct the phylogenetic analysis, the concatenated alignment of three data sets for the 72 genes from 3,654 species was used with six Rhodophyta as outgroups. There were a total of 44,187 positions for the matrix containing all codon positions (nt123), 29,458 positions for the matrix containing all but the third codon positions (nt12), and 14,724 amino acid

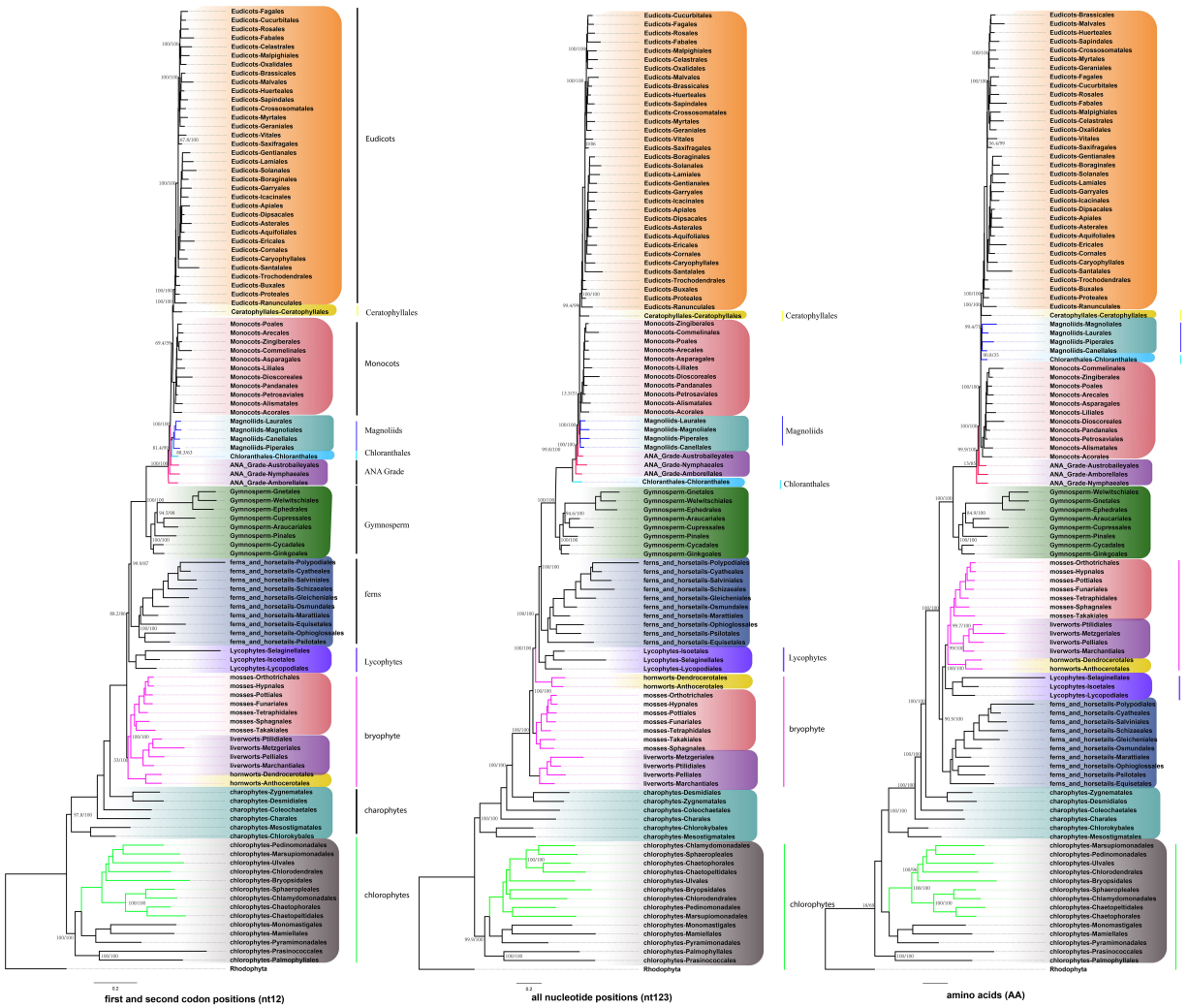


FIGURE 4 | Summary of the phylogenomic tree based on three data sets (nt12, nt123, and AA) of 72 plastid protein-coding genes of 3,654 green plants and six Rhodophyta using IQTREE. The colored branch and vertical lines (on the right side of the tree) represent the clade with conflicting phylogenetic placements based on three data sets. Totally, 631 taxa were obtained by selecting one to three representatives from each family and at least one taxon for the families with fewer taxon sampling, and the tree is represented at the order level in the figure.

(AA) positions. We used two programs: IQ-TREE and RAxML to construct the phylogenetic tree, but they both produced exactly the same topology (**Supplementary Figure 10**), so we only used IQ-TREE to illustrate our results (**Figure 3** and **Supplementary Table 4**). However, when we compared the phylogenetic clades using all the three data matrices (nt12, nt123, and AA) together, the phylogenetic discordance was observed for Chlorophyceae, Ceratophyllales, magnoliids, lycophytes, and bryophytes. The topologies are summarized in **Figures 4**, **5**, and the details of the phylogenetic trees are provided in **Supplementary Figures 4–8**.

There are two previous plastid-based phylogenetic analyses by Ruhfel et al. (2014) and Gitzendanner et al. (2018) where they used 360 and 1,879 taxa to study the green plants, respectively. In yet another study, by constructing a phylogenetic tree based on 80 genes along with 62 fossil calibration data, Li et al. (2019) predicted that the origin of crown angiosperms occurred in Upper Triassic, whereas other major angiosperms appeared during the Jurassic and Lower Cretaceous period. Recently, Li et al. (2021a) used 4,660 taxa comprising 433 families that nearly include all currently recognized families to produce a reliable relationship of flowering plants. Moreover, chloroplast genes have been extensively utilized to resolve taxonomical controversies of several plant lineages (Pryer et al., 2004; Sahu et al., 2015, 2016; Shen et al., 2017; Li et al., 2019, 2022; One Thousand Plant Transcriptomes Initiative, 2019). Although most topologies of our phylogenetic trees were consistent, there were some differences with the previous reports. For some debated clades, the phylogenetic trees were incongruent based on nt12, nt123, AA, and nuclear data set. The summary of the similarities and conflicts in topologies derived from these four data sets are presented in **Figure 5** and **Supplementary Table 4**.

All the phyla of green plants except charophytes was recovered as monophyletic. Within chlorophytes, the matrix nt12, nt123, and AA supported that Palmophyllales and Prasinococcales are the earliest-diverging lineage of the green plants (UFboot = 100%) (**Figure 5A**). Chlorophyceae is monophyletic and Ulvophyceae is a non-monophyletic group based on the matrix nt12, nt123, and AA. The matrix nt123 placed the Chlorophyceae as sister to other Ulvophyceae. The ASTRAL trees by both 1 KP (One Thousand Plant Transcriptomes Initiative, 2019), and Li et al. (2021b) supported the Chlorophyceae as sister to Ulvophyceae II (Bryopsidales) (**Figure 5B**). During the evolution of Streptophyta, charophyte lineages formed a paraphyletic assemblage with the land plants. Chlorokybales + Mesostigmatales are the earliest-branching lineage, and a clade of Zygnematales + Desmidiales is the sister group to the land plants, which is similar to the previous analyses, which includes the results from 1 KP (one thousand plant transcriptomes) project (Leliaert et al., 2012; Lemieux et al., 2016; Li et al., 2020).

Within Euphyllophyta, in the matrix nt12 and nt123, a well-supported Monilophyta was found to be a sister to Spermatophyta (UFboot = 100%), but the matrix AA indicated that Monilophyta is sister to bryophytes (UFboot = 100%). Within Monilophyta, matrix nt12 supported Ophioglossales as the earliest-diverging lineage (UFboot = 100%), while matrix nt123 supported Equisetales as the earliest branch (UFboot = 100%). A recent analysis of non-synonymous nucleotide data and translated amino acid data from 83 chloroplast genes across 30 taxa suggests that bryophytes are monophyletic (Sousa et al., 2020). Based on the AA analysis, Gitzendanner et al. (2018) recovered bryophyte clade as monophyletic. In our matrix AA analysis, we found bryophyte + lycophytes as sister to ferns (UFboot = 100%). With matrix nt123, hornworts, mosses, and liverworts were identified as the successive sister lineages of tracheophytes (UFboot = 100%). With matrix nt12, bryophytes were identified as monophyletic and positioned as sister to the vascular plants (**Figure 5D**), whereas 1KP also recovered extant bryophyte as monophyletic as per ASTRAL analysis based on the nuclear genes.

Both of these topologies were well supported by the previous research (Nickrent et al., 2000; Sugiura et al., 2004). It should be noted that the third codon position likely has a much faster rate of evolution and has reached the saturation level causing the variations in the phylogenetic tree (Simmons et al., 2006).

Within Spermatophyta, gymnosperms were designated as sister to angiosperms. Moreover, within gymnosperms, the subclades were well supported in all three data sets. The Cycadales + Ginkgoales clades were identified as sisters to the rest of the gymnosperms. The Gnetales, Welwitschiales along with Ephedrales, formed a clade (UFboot = 100%), which are sisters to the clade comprising Cupressales and Araucariales were

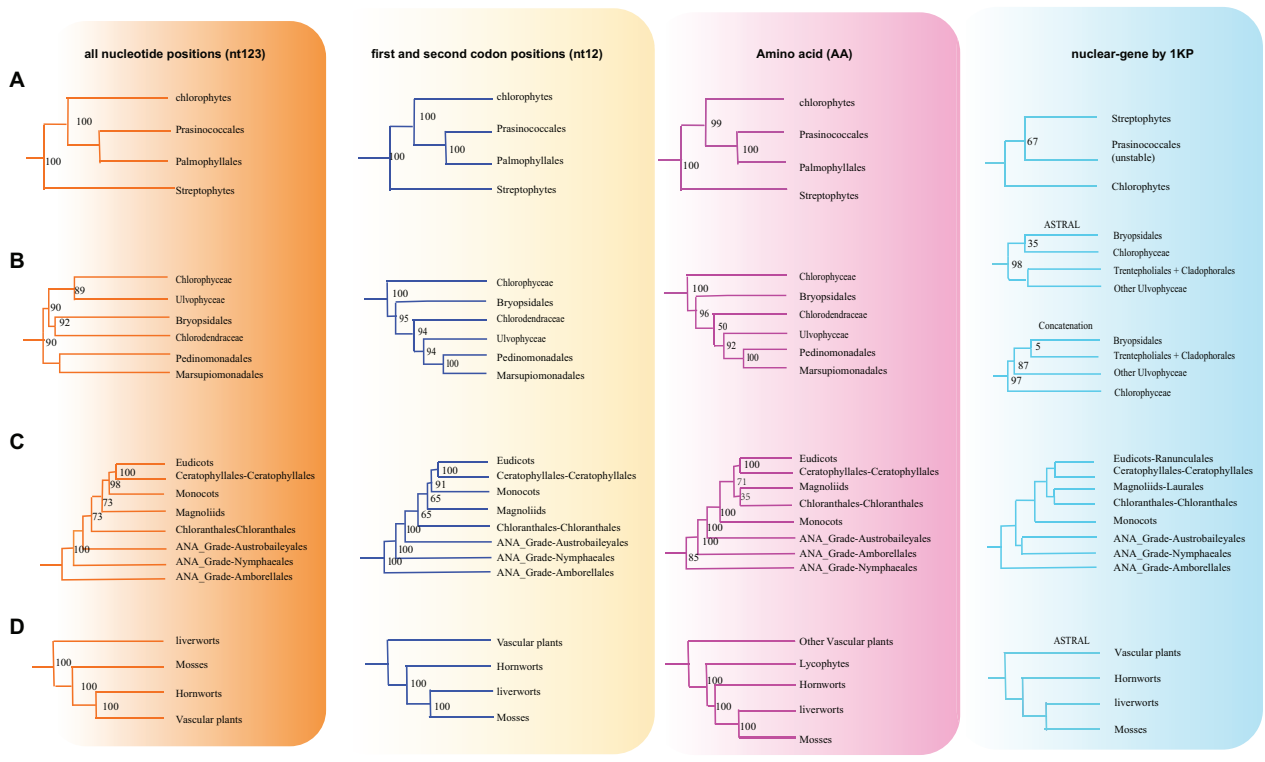


FIGURE 5 | Various branching orders for the phylogenetically discordant relationships. **(A)** Early Viridiplantae diversification, **(B)** early diversification of green algae, **(C)** the lineages of angiosperms, **(D)** early embryophyte diversification. The summarized topology is based on three data sets (nt12, nt123, and AA) of 72 protein-coding genes of 3,654 green plants and six Rhodophyta using IQTREE, including 1KP data set (nuclear gene-based).

not congruent with nuclear gene trees. In the 1KP project, the supermatrix of 410 single-copy nuclear gene family supports Gnetales as sister to Pinales, while coalescent analyses strongly support Gnetales sister to conifers (Araucariales, Cupressales and Pinales) (One Thousand Plant Transcriptomes Initiative, 2019).

Within angiosperms, in matrix nt12 and nt123, the Amborellales were recovered as the sister to all other angiosperms, followed by Nymphaeales. Nevertheless, Nymphaeales were placed as sisters to the remaining angiosperms based on the matrix AA (UFboot = 85%). Magnoliids were placed outside of the monocots in matrix nt123 and nt12 (UFboot = 100%), but based on the maxtrix AA, magnoliids and Chloranthales formed a sister clade to Ceratophyllales + eudicot (**Figure 5C**), which was consistent with the previous analyses (Guo et al., 2021). However, when we combined the data set from the study of Gitzendanner et al. (2018) with our AA sequences, magnoliids moved outside of the monocots (UFboot = 95%). Ruhfel et al. (2014) recovered Ceratophyllales as sister to the monocots using matrix nt12 with low support (*BS* = 52%). It should be noted that these discrepancies in tree topologies can be also attributable to biological phenomena like incomplete lineage sorting (ILS) and hybridization, as well as methodological challenges such as incorrect substitution model selection (Sousa et al., 2020; Yang et al., 2020; Guo et al., 2021). The relationship between COM clade supported Oxalidales as sister to Celastrales + Malpighiales. The major subclades were typically well supported in monocots and eudicots, but the position of Vitales, Gentianales, Petrosaviales, and Arecales remained uncertain. To further verify our phylogenetic analysis, the amino acid data from the study of Gitzendanner et al. (2018) were included, and the results showed that the species belonging to the same orders clustered together, and the topology of the major clade was consistent with the matrix nt12 (**Supplementary Figure 9**).

CONCLUSION

By performing a large-scale comparative analysis of 3,654 plastid genomes, we attempted to understand the evolution of plastome structure and gene content of green plants and revisited some long-standing uncertainties in green plant phylogeny. The structure of plastid genomes was mostly consistent in green plants and formed several gene blocks except in chlorophytes. We discovered that classes with similar functions likely constituted gene blocks. Some major genes such as the *psb* family probably coexisted in Viridiplantae and formed gene blocks. IR genes have doubled in size across terrestrial plants, and their GC content is substantially higher than that of non-IR genes. Regarding the green plant tree of life, more extensive taxon sampling indeed increased the phylogenetic resolution for some controversial clades. Our phylogenomic analyses have shown Chlorokybales + Mesostigmatales as the earliest branching lineages of streptophytes, and Zygnematales + Desmidiales were identified as the sister group of the embryophytes. In general, for some controversial clades that are deep within green plants, such as, bryophytes, dense taxon sampling did not improve phylogenetic accuracy anymore. Thus, to resolve the controversial deep-level clades, simply an increased taxon sampling may not be necessary or enough. In addition, plastid genome analysis alone seems unlikely to solve the relationship of these controversial clades (Ceratophyllales/Chloranthales). Using large numbers of nuclear genes or selecting the nuclear genes with stronger phylogenetic signals may help to answer these deep-level questions in the future studies.

MATERIALS AND METHODS

Taxon Sampling

We sampled 3,654 species including 3,648 representatives of green plants from 111 orders, 298 families, and six species of Rhodophyta as outgroups. The core chlorophyte clades, ferns and bryophytes, Mesangiospermae (comprising magnoliids, Chloranthales, monocots, Ceratophyllum, and Eudicots) were mainly focused in this study. We source our data from 3,246 published green plants plastid genomes from GenBank (as of January 18, 2019) and 731 previously generated plastomes from Ruili Botanical Garden (Liu et al., 2019). For multiple plastomes of the same taxon, we chose the plastome with a circular structure and a complete plastid genome. To make sure the high-quality data sets, we removed any species that had more than 50% gene missing in the same family. A total of six poorly annotated species (*Monoraphidium neglectum*, CM002678; *Nothoceros aenigmaticus*, NC-020259; *Nymphaea ampla*, NC-035680; *Allium sativum*, NC-031829; *Bambusa oldhamii*, NC-012927, and *Potentilla micrantha*, HG931056) were subjected to re-annotation with GeneWise v2.4.1 (Birney and Durbin, 2000). The complete list and the detailed information of 3,654 plastid genomes are provided in **Supplementary Table 1**.

Sequence Alignment

DNA sequences of protein-coding genes were extracted from each genome sequence according to the annotation files. Each protein-coding gene was processed individually with TranslatorX (Abascal et al., 2010) using MAFFT v7.310 (Katoh et al., 2002) to align the amino acid sequences and generated the corresponding nucleotide alignments, while poorly aligned positions were trimmed by TrimAl v1.1 (Capella-Gutiérrez et al., 2009) with the gappyout option. A total of seven genes: *ndhF*, *psaA*, *psaB*, *rpoB*, *rpoC1*, *rpoC2*, and *ycf2* had no information regarding gene annotation (Liu et al., 2019), and the genes with more than 50% missing alignment position were excluded from phylogenetic reconstruction. Both nucleotide and amino acid alignments of protein-coding genes were used for subsequent phylogenetic analyses.

Phylogeny and Gene Block Analyses

To evaluate the utility of the phylogenetic software, maximum likelihood (ML) analyses were both performed with IQ-TREE v1.6.10 (Nguyen et al., 2014) and RAxML v8.2.4 (Stamatakis, 2014). The best substitution models were identified based on the corrected Akaike information criterion (AICc) using ModelFinder embedded in IQ-TREE, and with 5,000 ultrafast

bootstrap (UFboot) replicates, together with GTR + F + R10 model for nucleotide sequences and JTT + F + R10 model for amino acid sequences.[1] ML analysis was also conducted using RAxML under the GTRCAT model for nucleotide and PROTGAMMAWAG model for amino acids, and the 100 bootstrap replicates were set to test the reliability of each node for RAxML.

The concatenated alignment comprising of 72 nucleotide genes was generated at the nucleotide level, and ML analyses were carried out using IQ-TREE with 5,000 UFboot replicates, together with GTR + F + R10 model. The coalescent analyses of 72 nucleotide genes were also preformatted and compared with the tree from concatenation analyses. Each gene tree was constructed using IQ-TREE with 5,000 UFboot replicates, but with best substitution model which was calculated by ModelFinder embedded in IQ-TREE. Based on the AICc, the species tree was detected from 72 gene trees by ASTRAL v4.11.1 (Mirarab et al., 2014).

To further evaluate the backbone relationships of the green plant's phylogeny, we assembled a smaller subset of 631 taxa derived from the complete taxon sampling. These 631 taxa were obtained by selecting one to three representatives from each family and at least one taxon for the families with fewer taxon sampling. The sequences of protein-coding genes were aligned and trimmed as above. ML analyses were only conducted with IQ-TREE under the partitioning scheme. The optimal partitioning schemes and best-fitting models of each scheme were determined with PartitionFinder v2 (Lanfear et al., 2012) based on AICc, and separate partitioning by gene was defined as the default.

To verify the topologies of the phylogenetic tree, the amino acid sequences of 72 genes of 1,901 samples in former research (Gitzendanner et al., 2018) were downloaded to analyze along with our data using the IQ-TREE. The Tree_doctor v1.3 (Hubisz et al., 2011) was used to obtain the simplified trees at order levels. The species of Rhodophyta was set as outgroups to re-root the result, and the iTOL[2] was used for data visualization.

Gene Block and Frequency Analyses

Based on transcript expression levels of plastid genes in *Arabidopsis*, the plastid genes are classified into eight clusters (Geimer et al., 2008). Although, the clustered genes likely belong to the same functional categories, whether these genes are also in the same position along the genome remains elusive. Therefore, we chose 1,517 complete ptDNA, compared the gene order in the same region of the ptDNA, and calculated the block frequency (**Supplementary Table 3**).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

[1]http://www.iqtree.org/doc/Substitution-Models

[2]https://itol.embl.de/

AUTHOR CONTRIBUTIONS

BZ and HL: conceptualization. SS and WM: data curation. TY, SS, and WM: formal analysis. XL and HL: funding acquisition and project administration. TY: investigation and visualization. TY and SS: methodology and writing—original draft. MS, BZ, and HL: supervision. TY, SS, YL, LY, MS, BZ, and HL: writing, reviewing, and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Key R&D Program of China (No. 2019YFC1711000), the National Natural Science Foundation of China (Nos. 32122010 and 31970229), the Shenzhen Municipal Government of China (No. JCYJ20170817145512476), the Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011), the NMPA Key Laboratory for the Rapid Testing Technology of Drugs, Collection of crop genetic resources research and application, BGI-Shenzhen, Shenzhen 518120, China (No. 2011A091000047), and Key Laboratory of Genomics, Ministry of Agriculture, BGI-Shenzhen, Shenzhen 518120, China, and Collaborative Innovation Center for Modern Crop Production co-sponsored by Province and Ministry. This study was a part of the 10KP project (https://db.cngb.org/10kp/). This work was also supported by China National GeneBank (CNGB; https://www.cngb.org/).

ACKNOWLEDGMENTS

We sincerely thank Susann Wicke for her helpful suggestions and inputs on an earlier draft of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.808156/full#supplementary-material

Supplementary Figure 1 | The gene constitution in the green plants. In the heat map, the data are displayed in a grid where each row represents order and each column represents average gene number in the order.

Supplementary Figure 2 | Overview of GC content in *psb* family.

Supplementary Figure 3 | Coding genes in IRs in Streptophyta. Coding genes in IRs and upstream are shown in blue and yellow, respectively.

Supplementary Figure 4 | Chloroplast phylogenomic tree based on the matrix nt123 of 72 protein-coding genes of 3,654 green plants and six Rhodophyta using IQTREE. The colors on the internal circle indicate different families, while the colors on the external circle indicate different orders.

Supplementary Figure 5 | Chloroplast phylogenomic tree based on the matrix aa of 72 protein-coding genes of 3,654 green plants and six Rhodophyta using

IQTREE. The colors on the internal circle indicate different families while the colors on the external circle indicate different orders.

Supplementary Figure 6 | Chloroplast phylogenomic tree based on the matrix nt12 of 72 protein-coding genes of 3,654 green plants using RaXML. The colors on the internal circle indicate different families while the colors on the external circle indicate different orders.

Supplementary Figure 7 | Chloroplast phylogenomic tree based on the matrix nt123 of 72 protein-coding genes of 3,654 green plants using RaXML. The colors in the internal circle indicate different families while the colors in the external circle indicate different orders.

Supplementary Figure 8 | Chloroplast phylogenomic tree based on the matrix aa of 72 protein-coding genes of 3,654 green plants and 1,901 species in the former research using IQTREE. The colors on the internal circle indicate different families while the colors on the external circle indicate different orders.

Supplementary Figure 9 | Summary of the phylogenomic tree based on matrix aa of 72 protein-coding genes of 3,654 green plants and 1,901 species obtained from earlier reports using IQTREE.

Supplementary Table 1 | The detailed information and characters of the species used in this study, including species information (species name, genus, family, order, data source), genome size, GC content, gene number, and intron number.

Supplementary Table 2 | The conserved gene order among Streptophyta. Different colors represent different genes.

Supplementary Table 3 | The gene blocks frequency in Streptophyta. The first column represents the gene block, and the second column is the number of times the gene block was observed in 1,517 ptDNA.

Supplementary Table 4 | Description of hexadecimal colors is used in **Supplementary Figures 6–8**.

REFERENCES

Abascal, F., Zardoya, R., and Telford, M. J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38, W7–W13. doi: 10.1093/nar/gkq291

Adachi, Y., Kuroda, H., Yukawa, Y., and Sugiura, M. (2011). Translation of partially overlapping psbD-psbC mRNAs in chloroplasts: the role of 5′-processing and translational coupling. *Nucleic Acids Res.* 40, 3152–3158. doi: 10.1093/nar/gkr1185

Arias-Agudelo, L. M., González, F., Isaza, J. P., Alzate, J. F., and Pabón-Mora, N. (2019). Plastome reduction and gene content in New World Pilostyles (Apodanthaceae) unveils high similarities to African and Australian congeners. *Mol. Phylogen. Evol.* 135, 193–202. doi: 10.1016/j.ympev.2019.03.014

Bellot, S., Renner, S. S., and Evolution. (2016). The plastomes of two species in the endoparasite genus Pilostyles (Apodanthaceae) each retain just five or six possibly functional genes. *Genome Biol.* 8, 189–201. doi: 10.1093/gbe/evv251

Birney, E., and Durbin, R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Res.* 10, 547–548. doi: 10.1101/gr.10.4.547

Braukmann, T., Kuzmina, M., and Stefanović, S. (2013). Plastid genome evolution across the genus Cuscuta (Convolvulaceae): two clades within subgenus Grammica exhibit extensive gene loss. *J. Exp. Bot.* 64, 977–989. doi: 10.1093/jxb/ers391

Brázda, V., Lýsek, J., Bartas, M., and Fojta, M. (2018). Complex Analyses of Short Inverted Repeats in All Sequenced Chloroplast DNAs. *BioMed Res. Int.* 2018, 1–10. doi: 10.1155/2018/1097018

Brouard, J.-S., Otis, C., Lemieux, C., and Turmel, M. (2010). The exceptionally large chloroplast genome of the green alga Floydiella terrestris illuminates the evolutionary history of the Chlorophyceae. *Genome Biol. Evol.* 2, 240–256. doi: 10.1093/gbe/evq014

Burleigh, J. G., and Mathews, S. (2004). Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *Am. J. Bot.* 91, 1599–1613. doi: 10.3732/ajb.91.10.1599

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Chen, S., Zhang, H., Wang, X., Zhang, Y., Ruan, G., and Ma, J. (2021). Analysis of Codon Usage Bias in the chloroplast genome of Helianthus annuus J-01. *IOP Conf. Series* 792:012009. doi: 10.1088/1755-1315/792/1/012009

Chen, X., Fang, D., Wu, C., Liu, B., Liu, Y., Sahu, S. K., et al. (2020). Comparative plastome analysis of root-and stem-feeding parasites of Santalales untangle the footprints of feeding mode and lifestyle transitions. *Genome Biol. Evol.* 12, 3663–3676. doi: 10.1093/gbe/evz271

Choi, I.-S., Jansen, R., and Ruhlman, T. (2019). Lost and found: return of the inverted repeat in the legume clade defined by its absence. *Genome Biol. Evol.* 11, 1321–1333. doi: 10.1093/gbe/evz076

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., et al. (2006). The complete chloroplast genome sequence of Pelargonium × hortorum: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol. Biol. Evol.* 23, 2175–2190. doi: 10.1093/molbev/msl089

Clark, M. S. (2013). *Plant molecular biology—a laboratory manual*. Berlin: Springer Science & Business Media.

Coenye, T., and Vandamme, P. (2005). Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. *FEMS Microbiol. Lett.* 242, 117–126. doi: 10.1016/j.femsle.2004.10.050

Cremen, M. C. M., Leliaert, F., Marcelino, V. R., and Verbruggen, H. (2018). Large diversity of nonstandard genes and dynamic evolution of chloroplast genomes in siphonous green algae (Bryopsidales, Chlorophyta). *Genome Biol.* 10, 1048–1061. doi: 10.1093/gbe/evy063

Eckardt, N. A. (2006). Genomic Hopscotch: Gene Transfer from Plastid to Nucleus. *Plant Cell* 18, 2865–2867. doi: 10.1105/tpc.106.049031

Fang, L., Leliaert, F., Zhang, Z.-H., Penny, D., and Zhong, B.-J. (2017). Evolution of the Chlorophyta: insights from chloroplast phylogenomic analyses. *J. Syst. Evol.* 55, 322–332. doi: 10.1111/jse.12248

Figueroa-Martinez, F., Jackson, C., and Reyes-Prieto, A. (2019). Plastid genomes from diverse glaucophyte genera reveal a largely conserved gene content and limited architectural diversity. *Genome Biol. Evol.* 11, 174–188. doi: 10.1093/gbe/evy268

Gao, L., Su, Y. J., and Wang, T. (2010). Plastid genome sequencing, comparative genomics, and phylogenomics: current status and prospects. *J. Syst. Evol.* 48, 77–93. doi: 10.1111/j.1759-6831.2010.00071.x

Geimer, S., Meurer, J., and Cho, W. K. (2008). Cluster Analysis and Comparison of Various Chloroplast Transcriptomes and Genes in Arabidopsis thaliana. *DNA Res.* 16, 31–44. doi: 10.1093/dnares/dsn031

Gitzendanner, M. A., Soltis, P. S., Wong, G. K. S., Ruhfel, B. R., and Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *Am. J. Bot.* 105, 291–301. doi: 10.1002/ajb2.1048

Guo, X., Fang, D., Sahu, S. K., Yang, S., Guang, X., Folk, R., et al. (2021). Chloranthus genome provides insights into the early diversification of angiosperms. *Nat. Commun.* 12:6930. doi: 10.1038/s41467-021-26922-4

Howe, C. J., Barbrook, A., Nisbet, R., Lockhart, P., and Larkum, A. (2008). The origin of plastids. *Philos. Transact. Royal Soc. London B* 363, 2675–2685.

Hubisz, M. J., Pollard, K. S., and Siepel, A. (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* 12, 41–51. doi: 10.1093/bib/bbq072

Jost, M., Naumann, J., Rocamundi, N., Cocucci, A. A., and Wanke, S. (2020). The first plastid genome of the Holoparasitic genus Prosopanche (Hydnoraceae). *Plants* 9:306. doi: 10.3390/plants9030306

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

Keller, J., Rousseau-Gueutin, M., Martin, G. E., Morice, J., Boutte, J., Coissac, E., et al. (2017). The evolutionary fate of the chloroplast and nuclear rps16 genes as revealed through the sequencing and comparative analyses of four novel legume chloroplast genomes from Lupinus. *DNA Res.* 24, 343–358. doi: 10.1093/dnares/dsx006

Kim, H. T., Chung, M. G., and Kim, K.-J. (2014). Chloroplast genome evolution in early diverged leptosporangiate ferns. *Molecules* 37:372. doi: 10.14348/molcells.2014.2296

Lanfear, R., Calcott, B., Ho, S. Y., and Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29, 1695–1701. doi: 10.1093/molbev/mss020

Lee, J.-H., and Manhart, J. R. (2002). Four embryophyte introns and psbB operon indicate Chlorokybus as a basal streptophyte lineage. *Algae* 17, 53–58. doi: 10.4490/algae.2002.17.1.053

One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Leliaert, F., Smith, D. R., Moreau, H., Herron, M. D., Verbruggen, H., Delwiche, C. F., et al. (2012). Phylogeny and molecular evolution of the green algae. *Crit. Rev. Plant Sci.* 31, 1–46.

Leliaert, F., Verbruggen, H., and Zechman, F. W. (2011). Into the deep: new discoveries at the base of the green plant phylogeny. *Bioessays* 33, 683–692. doi: 10.1002/bies.201100035

Lemieux, C., Otis, C., and Turmel, M. (2016). Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Front. Plant Sci.* 7:697. doi: 10.3389/fpls.2016.00697

Li, F.-W., Kuo, L.-Y., Pryer, K. M., Rothfels, C., and Evolution. (2016). Genes translocated into the plastid inverted repeat show decelerated substitution rates and elevated GC content. *Genome Biol.* 8, 2452–2458. doi: 10.1093/gbe/evw167

Li, H.-T., Luo, Y., Gan, L., Ma, P.-F., Gao, L.-M., Yang, J.-B., et al. (2021a). Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biol.* 19:232. doi: 10.1186/s12915-021-01166-2

Li, X., Hou, Z., Xu, C., Shi, X., Yang, L., Lewis, L. A., et al. (2021b). Large Phylogenomic Data sets Reveal Deep Relationships and Trait Evolution in Chlorophyte Green Algae. *Genome Biol. Evol.* 13:evab101. doi: 10.1093/gbe/evab101

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., et al. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* 5, 461–470. doi: 10.1038/s41477-019-0421-0

Li, L., Chen, X., Fang, D., Dong, S., Guo, X., Li, N., et al. (2022). Genomes shed light on the evolution of Begonia, a mega-diverse genus. *New Phytol.* 234, 295–310. doi: 10.1111/nph.17949

Li, L., Wang, S., Wang, H., Sahu, S. K., Marin, B., Li, H., et al. (2020). The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants. *Nat. Ecol. Evol.* 4, 1220–1231. doi: 10.1038/s41559-020-1221-7

Lin, C.-P., Huang, J.-P., Wu, C.-S., Hsu, C.-Y., and Chaw, S.-M. (2010). Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. *Genome Biol. Evol.* 2, 504–517. doi: 10.1093/gbe/evq036

Lin, C.-P., Wu, C.-S., Huang, Y.-Y., and Chaw, S.-M. (2012). The complete chloroplast genome of Ginkgo biloba reveals the mechanism of inverted repeat contraction. *Genome Biol. Evol.* 4, 374–381. doi: 10.1093/gbe/evs021

Liu, F., and Melton, J. T. III (2021). Chloroplast Genomes of the Green-Tide Forming Alga Ulva compressa: Comparative Chloroplast Genomics in the Genus Ulva (Ulvophyceae, Chlorophyta). *Front. Marin. Sci.* 8:668542. doi: 10.3389/fmars.2021.668542

Liu, H., Wei, J., Yang, T., Mu, W., Song, B., Yang, T., et al. (2019). Molecular digitization of a botanical garden: high-depth whole-genome sequencing of 689 vascular plant species from the Ruili Botanical Garden. *Gigascience* 8:giz007. doi: 10.1093/gigascience/giz007

Liu, Y., Cox, C. J., Wang, W., and Goffinet, B. (2014). Mitochondrial phylogenomics of early land plants: mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Syst. Biol.* 63, 862–878. doi: 10.1093/sysbio/syu049

Lu, Y., Ran, J.-H., Guo, D.-M., Yang, Z.-Y., and Wang, X.-Q. (2014). Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. *PLoS One* 9:e107679. doi: 10.1371/journal.pone.0107679

Lyko, P., and Wicke, S. (2021). Genomic reconfiguration in parasitic plants involves considerable gene losses alongside global genome size inflation and gene births. *Plant Physiol.* 186, 1412–1423. doi: 10.1093/plphys/kiab192

Marsolier-Kergoat, M.-C., and Yeramian, E. (2009). GC content and recombination: reassessing the causal effects for the Saccharomyces cerevisiae genome. *Genetics* 183, 31–38. doi: 10.1534/genetics.109.105049

Martín, M., and Sabater, B. (2010). Plastid ndh genes in plant evolution. *Plant Physiol. Biochem.* 48, 636–645. doi: 10.1016/j.plaphy.2010.04.009

Matsuo, M., Ito, Y., Yamauchi, R., and Obokata, J. (2005). The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast–nuclear DNA flux. *Plant Cell* 17, 665–675. doi: 10.1105/tpc.104.027706

Maul, J. E., Lilly, J. W., Cui, L., Miller, W., Harris, E. H., and Stern, D. B. (2002). The Chlamydomonas reinhardtii plastid chromosome: islands of genes in a sea of repeats. *Plant Cell* 14, 2659–2679. doi: 10.1105/tpc.006155

Meunier, J., and Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* 21, 984–990. doi: 10.1093/molbev/msh070

Millen, R. S., Olmstead, R. G., Adams, K. L., Palmer, J. D., Lao, N. T., Heggie, L., et al. (2001). Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* 13, 645–658. doi: 10.1105/tpc.13.3.645

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30, i541–i548. doi: 10.1093/bioinformatics/btu462

Mohanta, T. K., Mishra, A. K., Khan, A., Hashem, A., Abd_Allah, E. F., and Al-Harrasi, A. (2020). Gene loss and evolution of the plastome. *Genes* 11:1133 doi: 10.3390/genes11101133

Moore, M. J., Bell, C. D., Soltis, P. S., and Soltis, D. E. (2007). Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl. Acad. Sci.* 104, 19363–19368. doi: 10.1073/pnas.0708072104

Mower, J. P., Ma, P. F., Grewe, F., Taylor, A., Michael, T. P., VanBuren, R., et al. (2019). Lycophyte plastid genomics: extreme variation in GC, gene and intron content and multiple inversions between a direct and inverted orientation of the rRNA repeat. *New Phytol.* 222, 1061–1075. doi: 10.1111/nph.15650

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2014). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nickrent, D. L., Parkinson, C. L., Palmer, J. D., and Duff, R. J. (2000). Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* 17, 1885–1895. doi: 10.1093/oxfordjournals.molbev.a026290

Papanikolaou, N., Trachana, K., Theodosiou, T., Promponas, V. J., and Iliopoulos, I. (2009). Gene socialization: gene order, GC content and gene silencing in *Salmonella*. *BMC Genom.* 10:597. doi: 10.1186/1471-2164-10-597

Pryer, K. M., Schuettpelz, E., Wolf, P. G., Schneider, H., Smith, A. R., and Cranfill, R. (2004). Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. *Am. J. Bot.* 91, 1582–1598. doi: 10.3732/ajb.91.10.1582

Puttick, M. N., Morris, J. L., Williams, T. A., Cox, C. J., Edwards, D., Kenrick, P., et al. (2018). The interrelationships of land plants and the nature of the ancestral embryophyte. *Curr. Biol.* 28, 733.e–745.e. doi: 10.1016/j.cub.2018.01.063

Qu, X.-J., Fan, S.-J., Wicke, S., and Yi, T.-S. (2019). Plastome reduction in the only parasitic gymnosperm Parasitaxus is due to losses of photosynthesis but not housekeeping genes and apparently involves the secondary gain of a large inverted repeat. *Genome Biol. Evol.* 11, 2789–2796. doi: 10.1093/gbe/evz187

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., and Burleigh, J. G. (2014). From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* 14:23. doi: 10.1186/1471-2148-14-23

Ruhlman, T. A., Chang, W.-J., Chen, J. J., Huang, Y.-T., Chan, M.-T., Zhang, J., et al. (2015). NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biol.* 15:100. doi: 10.1186/s12870-015-0484-7

Sahu, S. K., Singh, R., and Kathiresan, K. (2015). Deciphering the taxonomical controversies of Rhizophora hybrids using AFLP, plastid and nuclear markers. *Aqu. Bot.* 125, 48–56. doi: 10.1016/j.aquabot.2015.05.002

Sahu, S. K., Singh, R., and Kathiresan, K. (2016). Multi-gene phylogenetic analysis reveals the multiple origin and evolution of mangrove physiological traits through exaptation. *Estuarine Coast. Shelf Sci.* 183, 41–51. doi: 10.1016/j.ecss.2016.10.021

Shaw, J., and Renzaglia, K. (2004). Phylogeny and diversification of bryophytes. *Am. J. Bot.* 91, 1557–1581. doi: 10.3732/ajb.91.10.1557

Shen, H., Jin, D., Shu, J.-P., Zhou, X.-L., Lei, M., Wei, R., et al. (2017). Large-scale phylogenomic analysis resolves a backbone phylogeny in ferns. *Gigascience* 7:gix116. doi: 10.1093/gigascience/gix116

Simmons, M. P., Zhang, L.-B., Webb, C. T., and Reeves, A. (2006). How can third codon positions outperform first and second codon positions in phylogenetic inference? An empirical example from the seed plants. *Syst. Biol.* 55, 245–258. doi: 10.1080/10635150500481473

Simpson, C. L., and Stern, D. B. (2002). The treasure trove of algal chloroplast genomes. Surprises in architecture and gene content, and their functional implications. *Plant Physiol.* 129, 957–966. doi: 10.1104/pp.010908

Sousa, F., Civáň, P., Foster, P. G., and Cox, C. J. (2020). The chloroplast land plant phylogeny: analyses employing better-fitting tree-and site-heterogeneous composition models. *Front. plant Sci.* 11:1062. doi: 10.3389/fpls.2020.01062

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

Strauss, S. H., Palmer, J. D., Howe, G. T., and Doerksen, A. H. (1988). Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proc. Natl. Acad. Sci.* 85, 3898–3902. doi: 10.1073/pnas.85.11.3898

Sugiura, C., Yamaguchi, K., Yoshinaga, K., Ueda, K., Yamada, K., Sugita, M., et al. (2004). Chloroplast Phylogeny Indicates that Bryophytes Are Monophyletic. *Mol. Biol. Evol.* 21, 1813–1819. doi: 10.1093/molbev/msh203

Sundararajan, A., Dukowic-Schulze, S., Kwicklis, M., Engstrom, K., Garcia, N., Oviedo, O. J., et al. (2016). Gene evolutionary trajectories and GC patterns driven by recombination in Zea mays. *Front. Plant Sci.* 7:1433. doi: 10.3389/fpls.2016.01433

Turmel, M., Otis, C., and Lemieux, C. (2009). The chloroplast genomes of the green algae Pedinomonas minor, Parachlorella kessleri, and Oocystis solitaria reveal a shared ancestry between the Pedinomonadales and Chlorellales. *Mol. Biol. Evol.* 26, 2317–2331. doi: 10.1093/molbev/msp138

Umate, P., Schwenkert, S., Karbat, I., Dal Bosco, C., Mlcòchová, L., Volz, S., et al. (2007). Deletion of PsbM in tobacco alters the QB site properties and the electron flow within photosystem II. *J. Biol. Chem.* 282, 9758–9767. doi: 10.1074/jbc.m608117200

Waltari, E., and Edwards, S. V. (2002). Evolutionary dynamics of intron size, genome size, and physiological correlates in archosaurs. *Am. Natural.* 160, 539–552. doi: 10.1086/342079

Wang, R.-J., Cheng, C.-L., Chang, C.-C., Wu, C.-L., Su, T.-M., and Chaw, S.-M. (2008). Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol. Biol.* 8:36. doi: 10.1186/1471-2148-8-36

Wicke, S., and Naumann, J. (2018). Molecular evolution of plastid genomes in parasitic flowering plants. *Adv. Bot. Res.* 85, 315–347. doi: 10.1016/bs.abr.2017.11.014

Wicke, S., Schneeweiss, G. M., Müller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297. doi: 10.1007/s11103-011-9762-4

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci.* 111, E4859–E4868. doi: 10.1073/pnas.1323926111

Wolf, P. G., Karol, K. G., Mandoli, D. F., Kuehl, J., Arumuganathan, K., Ellis, M. W., et al. (2005). The first complete chloroplast genome sequence of a lycophyte, Huperzia lucidula (Lycopodiaceae). *Gene* 350, 117–128. doi: 10.1016/j.gene.2005.01.018

Wu, C. S., and Chaw, S. M. (2014). Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. *Plant Biotechnol. J.* 12, 344–353. doi: 10.1111/pbi.12141

Wu, C.-S., and Chaw, S.-M. (2015). Evolutionary stasis in cycad plastomes and the first case of plastome GC-biased gene conversion. *Genome Biol. Evol.* 7, 2000–2009. doi: 10.1093/gbe/evv125

Wu, C.-S., Wang, Y.-N., Hsu, C.-Y., Lin, C.-P., and Chaw, S.-M. (2011). Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biol. Evol.* 3, 1284–1295. doi: 10.1093/gbe/evr095

Xiao-Ming, Z., Junrui, W., Li, F., Sha, L., Hongbo, P., Lan, Q., et al. (2017). Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. *Sci. Rep.* 7:1555. doi: 10.1038/s41598-017-01518-5

Xu, J.-H., Liu, Q., Hu, W., Wang, T., Xue, Q., and Messing, J. (2015). Dynamics of chloroplast genomes in green plants. *Genomics* 106, 221–231. doi: 10.1016/j.ygeno.2015.07.004

Yang, L., Su, D., Chang, X., Foster, C. S., Sun, L., Huang, C.-H., et al. (2020). Phylogenomic insights into deep phylogeny of angiosperms based on broad nuclear gene sampling. *Plant Commun.* 1:100027. doi: 10.1016/j.xplc.2020.100027

Zhang, D., Li, K., Gao, J., Liu, Y., and Gao, L.-Z. (2016). The complete plastid genome sequence of the wild rice Zizania latifolia and comparative chloroplast genomics of the rice tribe Oryzeae, Poaceae. *Front. Ecol. Evol.* 4:88. doi: 10.3389/fevo.2016.00088

Zhu, A., Guo, W., Gupta, S., Fan, W., and Mower, J. P. (2016). Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209, 1747–1756. doi: 10.1111/nph.13743

Conflict of Interest: TY, SS, YL, WM, XL, and HL were employed by the company Beijing Genomics Institute (BGI-Shenzhen).

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 13 April 2022
doi: 10.3389/fpls.2022.850521

Localized Phylogenetic Discordance Among Nuclear Loci Due to Incomplete Lineage Sorting and Introgression in the Family of Cotton and Cacao (Malvaceae)

Rebeca Hernández-Gutiérrez[1,2], Cássio van den Berg[3], Carolina Granados Mendoza[2], Marcia Peñafiel Cevallos[4], Efraín Freire M.[4], Emily Moriarty Lemmon[5], Alan R. Lemmon[6] and Susana Magallón[2]*

[1]Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México, Mexico City, Mexico, [2]Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3]Departamento de Ciencias Biológicas, Universidade Estadual de Feira de Santana, Feira de Santana, Brazil, [4]Herbario Nacional del Ecuador (QCNE), Instituto Nacional de Biodiversidad, Quito, Ecuador, [5]Department of Biological Science, Florida State University, Tallahassee, FL, United States, [6]Department of Scientific Computing, Florida State University, Tallahassee, FL, United States

OPEN ACCESS

Edited by:
Susann Wicke, Humboldt University of Berlin, Germany

Reviewed by:
John Gatesy, University of California, Riverside, United States
Agnes Scheunert, Bavarian Natural History Collections, Germany

***Correspondence:**
Rebeca Hernández-Gutiérrez
rebecahdezgtz@comunidad.unam.mx

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *07 January 2022*
Accepted: *14 March 2022*
Published: *13 April 2022*

Citation:
Hernández-Gutiérrez R, van den Berg C, Granados Mendoza C, Peñafiel Cevallos M, Freire M. E, Lemmon EM, Lemmon AR and Magallón S (2022) Localized Phylogenetic Discordance Among Nuclear Loci Due to Incomplete Lineage Sorting and Introgression in the Family of Cotton and Cacao (Malvaceae). Front. Plant Sci. 13:850521. doi: 10.3389/fpls.2022.850521

The economically important cotton and cacao family (Malvaceae *sensu* lato) have long been recognized as a monophyletic group. However, the relationships among some subfamilies are still unclear as discordant phylogenetic hypotheses keep arising when different sources of molecular data are analyzed. Phylogenetic discordance has previously been hypothesized to be the result of both introgression and incomplete lineage sorting (ILS), but the extent and source of discordance have not yet been evaluated in the context of loci derived from massive sequencing strategies and for a wide representation of the family. Furthermore, no formal methods have been applied to evaluate if the detected phylogenetic discordance among phylogenomic datasets influences phylogenetic dating estimates of the concordant relationships. The objective of this research was to generate a phylogenetic hypothesis of Malvaceae from nuclear genes, specifically we aimed to (1) investigate the presence of major discordance among hundreds of nuclear gene histories of Malvaceae; (2) evaluate the potential source of discordance; and (3) examine whether discordance and loci heterogeneity influence on time estimates of the origin and diversification of subfamilies. Our study is based on a comprehensive dataset representing 96 genera of the nine subfamilies and 268 nuclear loci. Both concatenated and coalescence-based approaches were followed for phylogenetic inference. Using branch lengths and topology, we located the placement of introgression events to directly evaluate whether discordance is due to introgression rather than ILS. To estimate divergence times, concordance and molecular rate were considered. We filtered loci based on congruence with the species tree and then obtained the molecular rate of each locus to distribute them into three different sets corresponding to shared molecular rate ranges. Bayesian dating was performed for each of the different sets of loci with the same parameters and calibrations. Phylogenomic discordance was detected between methods, as well as gene

histories. At deep coalescent times, we found discordance in the position of five subclades probably due to ILS and a relatively small proportion of introgression. Divergence time estimation with each set of loci generated overlapping clade ages, indicating that, even with different molecular rate and gene histories, calibrations generally provide a strong prior.

Keywords: gene tree congruence, Malvaceae, molecular heterogeneity, phylogenomic dating, phylogenetic discordance, species tree

INTRODUCTION

Deep, conflicting phylogenetic relationships are often found in angiosperm clades, and the advancement of molecular sequencing of large amounts of loci from different compartments, as well as the thorough application of phylogenetic and coalescence methods, have greatly contributed to solve some of them (e.g., Wang et al., 2019b; Koenen et al., 2020; Cai et al., 2021; Jost et al., 2021). Nevertheless, in some cases, even with numerous genes, phylogenetic relationships remain unsolved or poorly supported due to the high incongruence among gene histories and the obscuring signal of past evolutionary processes such as incomplete lineage sorting (ILS) and reticulation, e.g., in Amaranthaceae s.l. (Morales-Briones et al., 2021).

An important aspect to consider when using hundreds or thousands of loci is that molecular rate heterogeneity increases; thus, phylogenetic tree inference should consider incongruence and molecular rate heterogeneity (Dornburg et al., 2019). Phylogenies represent the basis for downstream evolutionary analyses, such as divergence time estimation, which uses molecular clock models that are sensitive to rate heterogeneity, biasing age estimates if rate heterogeneity is not considered appropriately (Angelis et al., 2018; Smith et al., 2018; Carruthers et al., 2020). The test of evolutionary hypotheses is hindered by the intricate phylogenetic relationships, which would show equivocal or unconclusive results on, for example, the origin and diversification of lineages, or ancestral state and biogeographic area reconstructions.

The family Malvaceae is the largest in the order Malvales, with 4,465 species and 245 genera (The Plant List, 2021) distributed in nine subfamilies, which comprise the traditional families Sterculiaceae, Tiliaceae, Bombacaceae, and Malvaceae *sensu stricto* (Alverson et al., 1999; Bayer et al., 1999). Many members of the family are important components of tropical ecosystems, and some others are of high economic importance (e.g., cotton, chocolate, cola nut, and durian). Malvaceae is highly diverse in growth forms, fruit types, floral morphology, and geographic and biome distribution. Understanding how this family evolved to reach such a high variation is a challenging task, starting from the phylogeny, since recalcitrant discordance in the relationships among some subfamilies, i.e., Helicterioideae, Sterculioideae, Tilioideae, Dombeyoideae, and Brownlowioideae (Alverson et al., 1999; Bayer et al., 1999; Nyffeler et al., 2005; Richardson et al., 2015; Hernández-Gutiérrez and Magallón, 2019; Cvetković et al., 2021), weakens the possible hypotheses about its evolution. The evolution of Malvaceae seems to be highly complex because nuclear genes show a different history from the plastome, but differences in the same genomic compartment are also present (Conover et al., 2019; Cvetković et al., 2021; Hernández-Gutiérrez et al., 2021; Wang et al., 2021). Importantly, there is no consensus on the relationships among some subfamilies due to conflicting, but highly supported resolutions, as observed in past but mostly in recent studies (Conover et al., 2019; Hernández-Gutiérrez and Magallón, 2019; Cvetković et al., 2021; Hernández-Gutiérrez et al., 2021; Wang et al., 2021).

Within Malvaceae, multiple whole-genome multiplications (WGM) have occurred, as observed by analyzing genomic data (Paterson et al., 2012; Wang et al., 2019a) but also inferred through chromosome counting (e.g., Costa et al., 2017). It has been hypothesized that deep reticulations gave rise to some major lineages of Malvaceae and that some of the resulting conflicting relationships are caused by ILS (Conover et al., 2019). The extent at which these two sources of phylogenetic discordance are causing of the contradictory hypotheses of Malvaceae phylogeny remains unknown. To analyze this question, numerous nuclear genes, and a larger taxon sampling, have the potential to inform about past processes underlying the intricate relationships among subfamilies of Malvaceae.

The timing of evolution of Malvaceae was previously estimated in a comprehensive study of the order Malvales, mostly based on plastid molecular markers (Hernández-Gutiérrez and Magallón, 2019). Although nuclear genes can potentially modify estimates of phylogenetic relationships, and phylogenomic data commonly violate molecular clock model assumptions, both factors consequently affect age estimates (Angelis et al., 2018). Because accurate divergence time estimations represent a framework to further analyze lineage evolution, here we examine to what extent gene conflict and molecular rate heterogeneity impact the divergence time estimation of Malvaceae. Using a comprehensive taxon sampling, our objective was to reconstruct the phylogenetic relationships of Malvaceae from nuclear genes. The specific aims of this study were to (1) investigate the presence of major phylogenetic discordance among hundreds of nuclear gene histories of Malvaceae; (2) evaluate the extent to which reticulation and ILS are causing discordance; and (3) to estimate divergence times considering discordance and heterogeneity in gene histories.

MATERIALS AND METHODS

Plant Material, Taxon Sampling, and DNA Extraction

DNA extraction was performed from silica dried tissue, as well as herbarium material (**Supplementary Table S1**), with a modified CTAB protocol (Doyle and Doyle, 1987) that includes

an additional treatment with RNAse A (Qiagen, Mexico City, Mexico) and proteinase K (recombinant, 1 mg/ml; Thermo Scientific, Mexico City, Mexico). The extraction and molecular procedures of Brazilian samples (**Supplementary Table S1**) were done at Laboratório de Sistemática Molecular de Plantas (LAMOL), Universidade Estadual de Feira de Santana. We included 96 species, each from a different genus, representing the nine subfamilies of Malvaceae s.l. (**Supplementary Table S1**). Nine species belonging to other families in the order Malvales were included as outgroups (**Supplementary Table S1**). To build a phylogenetic tree with a concatenated matrix, *Neurada procumbens* was selected for rooting the tree, following results obtained in a previous study (Hernández-Gutiérrez and Magallón, 2019). However, for rooting phylogenetic gene trees, different outgroups were selected because individual loci alignments have different taxon sampling due to sequencing capture variations (see details for each analysis below).

Plant Anchored Enrichment Strategies

Molecular data were generated through two target enrichment strategies in the Center for Anchored Phylogenomics at Florida State University.[1] Both strategies used the Angiosperm v.1 probe kit (Buddenhagen et al., 2016) which targets 499 nuclear exons that were found to be present in low or single copy in several species well distributed across the angiosperm phylogeny and 18 additional exons corresponding to selected selenium-tolerance genes. The rationale behind the design of this probe set is explained in detail by its authors (Buddenhagen et al., 2016), as well as in studies applying this kit to other angiosperm lineages (Lamiaceae: Fragoso-Martínez et al., 2017; Aristolochiaceae: Wanke et al., 2017). In general, the two strategies followed the same wet-lab procedures for library preparation, enrichment, and sequencing, which in summary were as follows. A Covaris E220 Focused-ultrasonicator was used to shear the DNA to a fragment size of 300–800 bp. A modification of the protocol of Meyer and Kircher (2010) was used to bind the adapters and indexes to the fragmented DNA with a Beckman-Coulter Biomek FXp liquid-handling robot. Indexed samples were pooled to carry out solution-based enrichment reactions with the Angiosperm v. 1 probe kit (Agilent Technologies Custom SureSelect XT kit), following manufacturer's protocol. Streptavidin coated magnetic beads were used to separate the enriched DNA fragments from the remaining genomic DNA. The enrichment strategies differ from each other in how indexes were assigned during the library preparation step. In the first strategy, each species was first linked to a unique index and then pooled with other species for enrichment, as it is conventionally done. In the second strategy, six distantly related angiosperm species (among them one species of Malvaceae) were first pooled and then assigned a single index prior to enrichment, a method called Anchored MetaPrep (Lemmon, 2015). In the present study, five control samples were processed with both enrichment strategies and incorporated in the phylogenetic analyses to cross-validate the use of both data sources. Enrichment reactions from both strategies were sequenced in one PE150 Illumina HiSeq 2500 lane at the Translational Science Laboratory in the College of Medicine at Florida State University, Tallahassee, Florida, United States.

Read Processing, Assembly, Orthology Assessment, and Alignment

All methods described in this section were performed in the Center for Anchored Phylogenomics. A detailed explanation of the bioinformatic methods employed can be found in Granados Mendoza et al. (2020), but in short, low-quality raw reads were filtered out with the CASAVA v. 1.8 pipeline using a high-chastity setting. Read demultiplexing was performed by ensuring perfect matches to one of 13 indexes developed in-house and reads with ambiguous matches were excluded. We used the method proposed by Rokyta et al. (2012) for read merging, because this method prevents merging at highly repetitive regions. Assembly followed the *quasi-de novo* strategy and used the Assembler.java program of Prum et al. (2015), with both merged and unmerged reads. The assembler first performs a divergent reference assembly, where reads are mapped to conserved regions of the target loci using three distantly related species to our target group (i.e., *Arabidopsis thaliana*, *Billbergia nutans*, and *Carex lurida*) that were included in the probe set design by Buddenhagen et al., 2016. Then, a second *de novo* assembly is carried out, where reads assembled in the first step serve as references to extend the assembly into the more variable flanking regions. Unambiguous base calls were assumed if no polymorphism was observed or if polymorphisms could be attributed to sequencing errors, assuming a binomial probability model with a probability of error = 0.1 and alpha = 0.05 (Buddenhagen et al., 2016). Heterozygous sites were coded following the IUPAC ambiguity codes, and if coverage was below 10, bases were called as N. To avoid cross contamination and inclusion of potential sequencing errors, assembled contigs with <30× mean coverage were excluded. Orthology assessment followed Prum et al. (2015) and was performed by grouping sequences by locus and calculating a distance matrix, where pairwise distances between two sequences corresponded to the percent of 20-mers found in both sequences. These distance matrices were then used to cluster sequences using the neighbor-joining algorithm (Saitou and Nei, 1987). If a single cluster was produced, we assumed no gene duplication for that specific locus. If more than one cluster was obtained, each cluster was considered as a different locus. Only clusters with more than 50% of the target species were used in further steps. MAFFT v.7.023b (Katoh and Standley, 2013) was used to generate preliminary alignments that were subsequently trimmed following Prum et al. (2015) and Hamilton et al. (2016). For trimming, an alignment site was considered as "good" when the most prevalent character state was shared across >50% of the sequences, then regions of 20 bp of each sequence were masked if they contained less than 15 "good" sites, and finally, sites having less than 56 unmasked bases were trimmed. The bioinformatic process

[1]www.anchoredphylogeny.com

of the data derived from the Anchored MetaPrep method follows Lemmon (2015). A total of 268 nuclear loci alignments were obtained after merging the information retrieved from both enrichment strategies (**Supplementary File 1**).

Concatenated Phylogeny

We aimed at constructing a phylogeny with a concatenated matrix. For this, we concatenated all loci in R (R Core Team, 2020) with the chopper package[2] and transformed this alignment to NEXUS format with the ips package (Heibl et al., 2019). To estimate the substitution model for each locus, we used PartitionFinder2 (Lanfear et al., 2016) implemented in the CIPRES Gateway (Miller et al., 2010), all models were evaluated with the "greedy" algorithm (Lanfear et al., 2012) and using RAxML (Stamatakis, 2014) for phylogenetic inference. The model GTR + I + G was identified as best-fitting for most of the loci. We conducted maximum likelihood (ML) inference with the concatenated matrix with RAxML v. 8.2.12 in the BEAGLE server from the Instituto de Biología of the National Autonomous University of Mexico (UNAM), using the partition sets that resulted from PartitionFinder2, tree search was set to 10. Bootstrap support for nodes was evaluated with 1,000 replicates. The nine species of the other Malvalean families were assigned as the outgroup.

To assess the support of individual loci in relation to each node, we took the ML topology and generated reverse constraints for the 110 nodes of the tree. Then, heuristic searches for each constraint and an unconstrained topology were performed with maximum parsimony (MP), followed by the inclusion of each locus individually, in order to assess their relative contribution to each node, in an analogous fashion to Lee et al. (2011). The resulting logs were processed with TreeRot v. 3 (Sorenson and Franzosa, 2007) to generate trees with individual values for each locus, and with a custom python script we extracted data from the trees with all the values (loci/nodes). With this data, we calculated for each locus: (1) number of nodes with positive values (supporting locus), (2) number of nodes with negative values (conflicting locus), (3) positive–negative, and (4) sum of all individual scores. For the nodes, we calculated (1) number of loci with positive values, (2) number of loci with negative values, (3) positive–negative, and (4) sum of individual scores, which corresponds to the overall Bremer support for that node.

Species Tree Estimation

We performed a site-based analysis (i.e., without *a priori* specification of gene trees) with the concatenated matrix to estimate the species tree under the multispecies coalescent model (MSC) conducted in SVDquartets (Chifman and Kubatko, 2014) implemented in PAUP* v.4.0a166 (Swofford, 2002). The evaluation was performed for a maximum of 100,000 random quartets and statistical support for nodes was assessed by the calculation of 1,000 bootstrap replicates.

[2]https://github.com/fmichonneau/chopper

A summary coalescence method was also implemented. For this, we first estimated phylogenetic trees for each locus with maximum likelihood in RAxML v.8.2.12 (Stamatakis, 2014), setting GTR+G as the substitution model, 100 tree searches, and 1,000 bootstrap replicates. Bifurcations with bootstrap support ≤20 were collapsed with the program nw_ed of Newick Utilities v.1.6 (Junier and Zdobnov, 2010). A file containing the gene trees with low supported branches collapsed was the input for ASTRAL-III v.5.7.3 (Zhang et al., 2018). The support for branches was evaluated with local posterior probability (LPP).

Phylogenetic Discordance Source

To explicitly evaluate the extent to which reticulation and ILS are causing phylogenetic discordance we used QuIBL (Quantifying Introgression *via* Branch Lengths; Edelman et al., 2019). For each triplet of species, QuIBL extracts the frequency of topologies formed by that triplet in all gene trees. Each triplet topology has one internal branch (considering one and the same outgroup for all the triplets) and QuIBL calculates the likelihood of two distribution models of the length of this branch. One model considers that the branch length derives from a proportion of ILS only, and the second model considers ILS plus the proportion of introgressed loci. Both models are examined for each triplet and are evaluated with Bayesian Information Criterion (BIC). In this study, a reduced taxon sampling was used because (1) this analysis requires that all species are present in every gene tree, and (2) we wanted to evaluate the discordance at a deep phylogenetic level, i.e., at the divergence of subfamilies. To reduce our taxon sampling, we used the R package treeplyr v.0.1.10 (Uyeda and Harmon, 2020) to prune the trees corresponding to the selected sampling of species. Thus, for this analysis, we used 123 gene trees from RAxML, each tree with 18 species representing the nine subfamilies and one species as the outgroup for all the triplets (*Muntingia calabura*). The ASTRAL tree was used for interpreting QuIBL results by distinguishing topologies that were discordant from those that resembled this species tree.

Divergence Time Estimation

Molecular dating based on genomic data (i.e., hundreds or thousands of genes) may be challenging, as gene histories and molecular rate could be highly heterogeneous (Carruthers et al., 2020). This heterogeneity produces two general issues in molecular dating. One of them is the usual violation of the molecular clock model, exacerbated as more data are included, making it difficult to obtain accurate estimates (Smith et al., 2018; Carruthers et al., 2020). One solution to this issue is the "gene shopping" approach (Smith et al., 2018), where genes or loci are selected if they behave in a more clock-like fashion, with respect to other loci. The other issue is that applying one clock model to a large dataset may yield wrong estimates due to high substitution rate heterogeneity (Angelis et al., 2018; Nie et al., 2020), which may be solved by partitioning the data set in different clock regimes (Nie et al., 2020).

Here, we aimed to identify the extent of rate heterogeneity in our molecular dataset and whether this impacts age estimates. For this, we applied a combination of approaches to overcome gene history conflict and both heterogeneity issues, first by dividing the complete loci dataset in sets of loci that differ in substitution rate variation (attending the molecular clock issue) and by applying different clock models to each of these sets of loci (addressing the issue of one model fitting high heterogeneity). We compared the results among three sets of loci that differ in rate variance, additionally comparing a fourth analysis with the concatenated dataset but partitioned by the three sets of loci, and a fifth analysis of few loci with low rate variance. The next sections describe the filtering of loci and analyses.

"Gene Shopping": Data Filtering

First, we used SortaDate (Smith et al., 2018) scripts to sort gene molecular behavior, following a "gene shopping" framework (Smith et al., 2018). SortaDate scripts were implemented in python 2.7 and it was used along with the software phyx (Brown et al., 2017) to select those loci that shared similar rate variation. The input files were the individual, rooted gene trees, which we obtained from the RAxML analyses described above (268 trees), and the rooted species tree, which was the ASTRAL species tree because it is fully resolved. Species and gene trees were rooted with the pxrr function from the phyx software (Brown et al., 2017). We sorted the trees based on the proportion of bipartitions shared with the species tree, then by the root-to-tip variance, and lastly by tree length.

From the results of SortaDate, we set the arbitrary criterium to select those trees that had at least 0.3 proportion of bipartitions corresponding to the species tree, which resulted in 123 gene trees. From this set, we calculated terciles from the root-to-tip variance and obtained three sets of 41 trees each. Thus, the first tercile has a low variance and the third tercile the highest variance. Note that the molecular rate variance was not necessarily related to the proportion of bipartitions (i.e., gene tree discordance). The sequence alignments of individual loci corresponding to the selected sets of trees were then concatenated to perform dating analyses (three molecular matrices each with 41 loci). Additionally, we wanted to analyze if applying a molecular clock model to different partitions affects the estimates, so we concatenated the three sets of loci, obtaining a molecular matrix with 123 loci with three partitions. Moreover, to examine if the homogeneity and number of loci affect the estimates, we selected five loci corresponding to those that had the lowest rate variance (i.e., closer to a strict clock fashion) and built a fifth molecular matrix.

Dating Analyses

We estimated divergence times with BEAST2 v.2.6.3 (Bouckaert et al., 2019). We performed five dating analyses: one for each set of 41 concatenated loci, one for loci of all three sets partitioned by set, and another with five loci with the lowest molecular rate variance to evaluate whether estimates are affected when using the least heterogeneous molecular dataset, that is, fitting to a single clock regime ("clock-likeness" approach). We applied the following settings to the five analyses. In BEAUti v.2.6.3, we implemented a GTR + G molecular substitution model, using empirical base frequencies, molecular clock set as uncorrelated with rates obtained from a log-normal prior distribution (UCLN; Drummond et al., 2006) and a birth-death tree prior. We constrained the topology to resemble the analysis with ASTRAL only for the highly supported subfamilies and major clades (i.e., all subfamilies belonging to a major group), but left unconstrained the relationships among and inside these clades.

We constrained subfamilies to be monophyletic, this excluded Helicterioideae and Byttnerioideae; relationships within subfamilies were not constrained. We applied a secondary calibration to the root of the tree, i.e., the crown node of Malvales, as a uniform prior distribution with minimum value of 110.48 Ma and maximum value of 138.33 Ma, as obtained from the BEAST analysis performed by Ramírez-Barahona et al. (2020). Eight calibrations informed by the fossil record (**Supplementary Table S2**) were applied to the crown group of Malvaceae and to most of the subfamilies. To set the calibrations, we used uniform distributions with the minimum value being the upper bound of the stratigraphic epoch of each fossil, and the maximum value 138.33 Ma (the maximum value assigned to the root). We ran two independent analyses with 500–600 million generations each, sampling parameters every 5,000 steps. The analyses in BEAST2 were performed in the server BEAGLE of Instituto de Biología (UNAM). For each analysis, the resulting estimates were summarized in LogCombiner v.2.6.3, removing 20% of the samples as burn-in of the posterior parameter values, and 70% of the posterior sampling of trees. The Maximum Clade Credibility (MCC) tree and node mean heights were obtained in TreeAnnotator v.2.6.3. The MCC tree of each analysis was visualized in FigTree v.1.4.4[3] and their annotated data were extracted with the R package treeio (Wang et al., 2020) for comparison among the four different analyses. Finally, we tested whether the prior settings were constraining the estimates by running an analysis without considering a molecular dataset and only including the prior specifications.

RESULTS

Taxon and Genetic Sampling

In this study, 96 species of Malvaceae and nine outgroup species representing other families of Malvales were considered (**Supplementary Table S1**). By integrating the results from the two enrichment strategies, we obtained 268 potentially single-copy nuclear loci. From the complete 268 gene sampling, 28 were captured only through the conventional AHE method, whereas the rest were captured with both techniques (**Supplementary Table S1**).

[3]github.com/rambaut/figtree

Concatenated Phylogeny

In the concatenated dataset of all 268 nuclear loci, the representatives of the nine non-Malvaceae families were designated as outgroup, but the relationships among them were mostly weakly supported (**Supplementary Figure S1**). The only highly supported (100 Bootstrap support, BS) relationship was between *M. calabura* (Muntingiaceae) and *Bdallophytum americanum* (Cytinaceae). Within Malvaceae, two clades are recovered, Byttneriina and Malvadendrina. Byttneriina comprises two monophyletic subfamilies, Grewioideae and Byttnerioideae. Grewioideae is strongly supported, as well as the relationships within it, whereas Byttnerioideae is moderately supported (72 BS) as a monophyletic group. The rest of the subfamilies are included in Malvadendrina (**Figure 1**). Most members of Helicterioideae (except *Durio zibethinus*) form a clade that is the sister group of the rest of the subfamilies, which form two groups. One group comprises *D. zibethinus* as the sister taxon of a group formed by the monophyletic, highly supported subfamilies Sterculioideae, Tilioideae, and Brownlowioideae + Dombeyoideae. The other group is Malvatheca (**Figure 1**), where *Chiranthodendron pentadactylon* is the sister of the remaining members of the group, and *Ochroma pyramidale* is the sister taxon of Bombacoideae + Malvoideae (**Supplementary Figure S1**).

We retrieved relatively high Bremer support for all the loci, indicating low conflict (**Supplementary Table S3**). Of all the loci, only one presented more conflicting than supported nodes (L256), but its overall sum of supports is positive. On the other hand, two loci presented more supported than conflicting nodes (L149 and L129) but with an overall negative sum of supports. All other loci have support for most of the nodes, and the fact that the overall sums of support values for either loci or nodes are always positive, indicates that even when there are some negative values, these are of smaller magnitude in relation to the positive support for all loci.

Species Tree Estimation

The results obtained with SVDquartets (**Figure 1A**) yielded many weakly supported bipartitions due to discordance in the bootstrap replicates. Highly supported (85–100 BS) clades were (1) Malvaceae

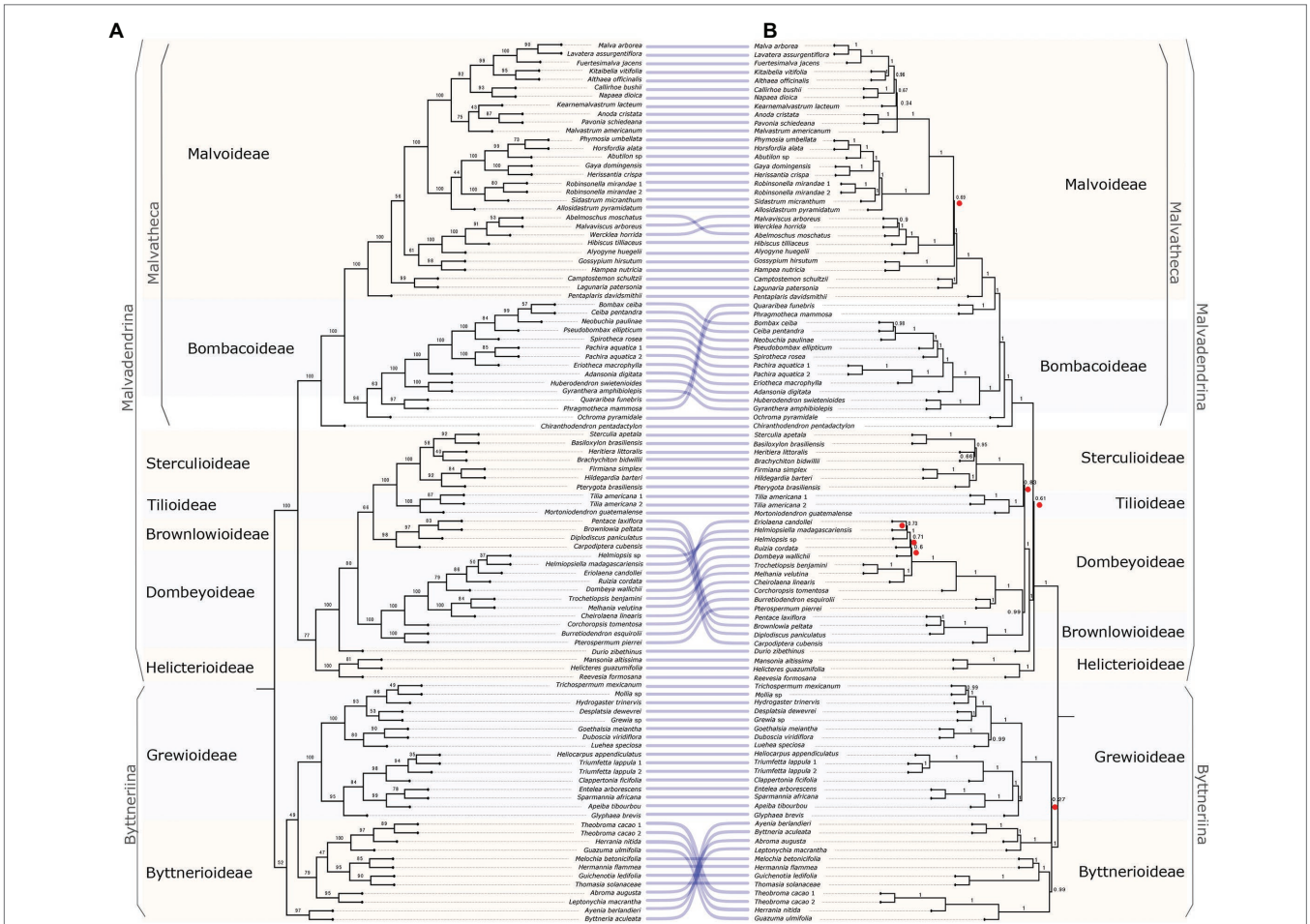


FIGURE 1 | Species trees of Malvaceae derived from two coalescence methods. **(A)** Species tree from SVDquartets. Numbers associated to nodes represent Bootstrap values. **(B)** Species tree from ASTRAL. Numbers associated to nodes represent local posterior probabilities (LPP). Red circles indicate relationships with low quartet score (<40%). To visualize similarities between the two analysis, purple lines connect species between trees.

as a whole; (2) Grewioideae; and (3) a clade containing *Durio zibethinus* as sister to Dombeyoideae, Brownlowioideae, Tilioideae, and Sterculioideae. These subfamilies are strongly supported as monophyletic, but the relationships among them are poorly supported (**Figure 1A**). Another group of highly supported relationships include (1) the Malvatheca clade; (2) *Chiranthodendron* as the sister lineage of the rest of Malvatheca species; (3) *Ochroma* as sister to *Quararibea* + *Phragmotheca* and Bombacoideae; and (4) Malvoideae.

The resulting species tree from the ASTRAL analysis (**Figure 1B**) shows strongly supported clades (1 LPP), such as Byttneriina and Malvadendrina. Within Byttneriina, highly supported clades are Grewioideae and part of Byttnerioideae (i.e., excluding tribe Byttnerieae, here comprising *Leptonychia*, *Byttneria,* and associated genera). The relationship of Helicterioideae (excluding *Durio*) and the rest of Malvadendrina are poorly supported (0.61 LPP). Within Malvadendrina, *Durio* is strongly supported as the sister of a clade comprising the subfamilies Sterculioideae, Tilioideae, Brownlowioideae, and Dombeyoideae. Brownlowioideae and Dombeyoideae are highly supported (1 LPP) as sister clades, and Sterculioideae and Tilioideae are moderately supported (0.83 LPP). Strongly supported relationships within Malvatheca are the placement of *Chiranthodendron* as the sister to the remaining species of Malvatheca, and successively *Ochroma* as the sister to the remaining species. *Quararibea* + *Phragmotheca* forms a clade that is sister to the Malvoideae.

The normalized quartet score (QT), which is the proportion of quartets in gene trees concordant with the species tree, is 0.9, meaning that concordance among gene trees is of 90% for the entire phylogeny. However, there are branches with low QT (<40%), indicating high gene tree discordance, coinciding with short branches (**Figure 1B**) and places of incongruent relationships with the SVDquartets tree (**Figure 1A**). Some of these branches are Byttnerioideae and some members of this subfamily that form a clade sister to Grewioideae; Malvatheca and its relationship with the rest of Malvadendrina; and branches within the subfamilies Dombeyoideae and Malvoideae.

We detected phylogenetic discordance by examining the low SVDquartets bootstrap values (**Figure 1A**) and the quartet score (QT) from the ASTRAL analysis (**Figure 1B**). Places with high discordance are the relationship between Helicterioideae and the four subfamilies Dombeyoideae, Tilioideae, Brownlowioideae, and Sterculioideae (77 BS; **Figure 1**); and Helicterioideae and the rest of Malvadendrina (37.51 QT; **Figure 1**); the relationship between Brownlowioideae and Sterculioideae + Tilioideae (66 BS; **Figure 1**); and Brownlowioideae and Dombeyoideae (42.45 QT; **Figure 1**). Byttnerioideae appeared as monophyletic in the analysis with a concatenated matrix (**Supplementary Figure S1**), but paraphyletic in the rest of the analyses (SVDquartets and ASTRAL; **Figure 1**), as well as in the temporally calibrated trees.

Phylogenetic Discordance Source

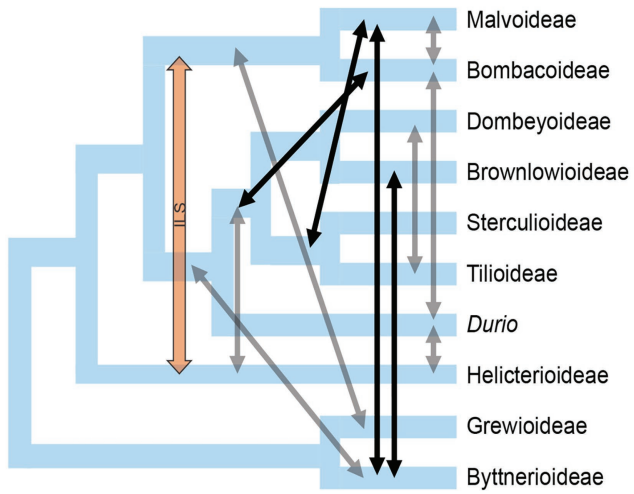
We evaluated the proportion of ILS and introgression in the discordant gene trees with QuIBL (Edelman et al., 2019), a method that analyzes triplet topologies present in the gene trees. QuIBL extracts branch lengths in each triplet topology to test two models of branch length distribution: one model includes a distribution generated only by ILS, and the other includes two distributions, one for ILS only and another for introgression. Model selection was obtained with BIC values, selecting those values that were significantly different with dBIC <−10 or >10, as recommended by Edelman et al. (2019). We examined the discordance of the relationships among subfamilies by including two representatives of each subfamily and a sample of 108 gene trees (**Supplementary Table S4**), resulting in 816 triplets and 2,248 topologies. **Table 1** summarizes QuIBL results that showed significant values (see **Supplementary Table S4** for detailed, significant results). Significant results suggest that 62 discordant topologies are caused by introgression, and three are caused by ILS (**Supplementary Table S4**). We summarized QuIBL results considering that some topologies represent a single introgression event, for example, triplets that have different species but of the same subfamily have equal values, thus corresponding to a single introgression event that is ancestral to the divergence of the species included. This contrast with the results found between Byttnerioideae and Malvoideae, where two different genera yielded different proportions of introgression (**Figure 2**). We obtained high proportions of ILS across the phylogeny of Malvaceae (**Figure 2**; **Table 1**), but according with the preferred model, the discordance can only be explained jointly with introgression given that 0.9–3.7% of the loci are introgressed (**Table 1**). We identified 12 main events of introgression that involve all nine subfamilies, and one event of ILS alone (i.e., without introgression) in Helicterioideae-Malvatheca (**Table 1**; **Figure 2**).

The proportion of introgressed loci is relatively high between the following pairs: Byttnerioideae-Brownlowioideae; Bombacoideae-(Dombeyoideae + Brownlowioideae + Tilioideae)-; Byttnerioideae-Malvoideae; and Malvoideae-(Sterculioideae + Tilioideae; **Figure 2**). Relatively low introgression is observed between Dombeyoideae-Tilioideae; Byttnerioideae-(Dombeyoideae + Sterculioideae + Tilioideae + *Durio*); Malvoideae-Bombacoideae; Bombacoideae-*Durio*; and Helicterioideae-*Durio* (**Figure 2**; **Table 1**). Given that the species tree (ASTRAL and SVDquartets) shows that *Durio* is separated from the rest of Helicterioideae, we describe the QuIBL results distinguishing Helicterioideae, with *Reevesia* as representative, from *Durio*. Introgression between Helicterioideae and the subfamilies Brownlowioideae, Dombeyoideae, Sterculioideae, and Tilioideae has the same magnitude, it is accompanied by a high proportion of ILS (96–97% of loci show ILS; **Supplementary Table S4**), and the tree counts are relatively similar among the three possible topologies, all of which indicate that ILS is highly frequent among these groups, but the signal is obscured due to a low but significant proportion of introgression. In turn, the trees with Helicterioideae as sister to Malvatheca are probably due to ILS only, and not introgression.

Divergence Time Estimation

To know whether the heterogeneity of molecular substitution rate, characteristic of genomic data, affects the estimation of divergence times, we conducted five dating analysis. First, we followed a

A



B

Introgressed pair groups	Total non-ILS prop. (%)
Byttnerioideae-Brownlowioideae	**3.70**
Bombacoideae-Dombeyoideae+Brownlowioideae+Tilioideae	**2.78**
Byttnerioideae-Malvoideae	**1.85**
Malvoideae-Sterculioideae+Tilioideae	**1.85**
Byttnerioideae-Dombeyoideae+*Durio*+Sterculioideae+Tilioideae	0.93
Helicterioideae-*Durio*	0.93
Bombacoideae-*Durio*	0.93
Dombeyoideae-Tilioideae	0.93
Helicterioideae-Brownlowioideae+Sterculoideae+Tilioideae+Dombeyoideae	0.93
Grewioideae-Malvoideae+Bombacoideae	0.93
Malvoideae-Bombacoideae	0.93
Byttnerioideae-Malvoideae	0.93

FIGURE 2 | Summary of the sources of phylogenetic discordance obtained from QuIBL. **(A)** Species tree with the relationships among Malvaceae subfamilies derived from ASTRAL. Arrows indicate the direction of introgression or ILS events: black arrows represent relative strong introgression (>1% total non-ILS proportion), gray arrows represent relative weak introgression (<1% total non-ILS proportion), and orange arrow represents ILS. **(B)** Total proportion (%) of loci that show introgression between pairs of subfamilies or groups of subfamilies; relative strong introgression (>1%) shown in bold. See **Supplementary Table S4** for detailed results.

"gene shopping" approach to filter loci from the complete 268 loci sampling. Loci were selected first by the proportion of splits (bipartitions) according to the species tree from ASTRAL, and then by molecular rate variance, resulting in 123 loci. We found heterogeneity in the molecular rate variance (**Figure 3A**; **Supplementary Table S5**); thus, thresholds were applied to obtain three sets, each one including 41 different loci sharing relatively similar molecular rate variance (**Figure 3A**). We performed two additional analyses, one with the three concatenated sets (123 loci) and another analysis with five concatenated loci that had the lowest rate variances to test whether number of loci and lower rate heterogeneity are influencing age estimates.

We tested if the priors were constraining the estimates instead of being informed by the molecular datasets, found that the molecular datasets are informing the posterior density (**Supplementary Figure S5**). In general, age estimates for clades are similar among the five different sets and their 95% Highest Posterior Density (HPD) intervals overlap (**Figure 3B**; **Supplementary Table S6**; **Supplementary File 2**). This result is not maintained, however, when some phylogenetic relationships are different, for example, Helicterioideae is sister to Malvatheca or to the rest of Malvadendrina in the different analyses, so its age varies the most (**Figure 3B**; **Supplementary Table S6**). In general, we note that the set with the highest molecular rate variances (set3) yielded older ages (**Figure 3B**), but the difference between the sets with low and medium rate variance (set1 and set2, respectively) was not pronounced (**Figure 3A**). The length of the HPD intervals is similar among the five analyses when considering the major clades in Malvaceae, *ca.* 19.8–23.8 million years for crown age and *ca.* 16.7–21.1 million years for stem age.

Considering that the five analyses yielded overlapping estimates and that the concatenated dataset of the three sets (concat_3sets) overall generated narrower 95% HPD intervals (**Supplementary Table S6**), that is, more precise estimates, we present the results of this dataset. Our results indicate an origin (stem age) of Malvaceae with a mean age of 126.5 Ma (Million years ago; 134–118 Ma 95% HPD; **Figure 4**), and a diversification age (crown age) with a mean of 107.71 Ma (114–100 Ma 95% HPD; **Figure 4**), both in the Lower Cretaceous. The nine subfamilies originated in the Upper Cretaceous, between 98.9 and 77.5 Ma (**Supplementary Table S6**), and diversified between the Upper Cretaceous and early Paleogene (74–59 Ma; **Figure 4**), except Helicterioideae and Tilioideae, which diversified in the early Eocene (56–47 Ma; **Figure 4**).

DISCUSSION

Phylogenetic Relationships in Light of ILS and Introgression

Since the circumscription of Malvaceae s.l., the motivation for resolving its phylogenetic relationships has been to investigate intriguing aspects of the family's evolution, such as its biogeographic distribution, paleontological evidence, or life history traits (Alverson et al., 1998, and references therein). More than 20 years later, the same motivation remains, and some key questions regarding Malvaceae evolution are still difficult to trace mostly due to conflicting phylogenetic results (e.g., Conover et al., 2019; Hernández-Gutiérrez et al., 2021). Consequently, in each independent study, where phylogenetic relationships are inferred *de novo,* new evolutionary hypotheses are formulated, instead of having a hypothesis that includes discordance sources in the evolution of Malvaceae and based on a consensus on the relationships within the family. Here,

TABLE 1 | Source of phylogenetic discordance due to introgression and ILS between pairs of taxa.

Taxa pairs	Subfamily groups	ILS proportion	Non-ILS proportion	BIC2	BIC1	dBIC	Total non-ILS prop. (%)
Guichenotia-Carpodiptera, Guichenotia-Brownlowia, Theobroma-Carpodiptera, Theobroma-Brownlowia	Byttnerioideae-Brownlowioideae	0.00	1.00	−30.91	−18.65	−12.26	3.70
Brownlowia-Pachira, Carpodiptera-Pachira, Corchoropsis-Pachira, Cheirolaena-Pachira, Mortoniodendron-Pachira	Bombacoideae-Brownlowioideae+ Dombeyoideae+Tilioideae	0.00	1.00	−33.99	−21.12	−12.87	2.78
Guichenotia-Pentaplaris	Byttnerioideae-Malvoideae	0.75	0.25	−45.91	−35.53	−10.37	1.85
Heritiera-Pentaplaris, Brachychiton-Pentaplaris	Sterculioideae-Malvoideae	0.50	0.50	−34.70	−22.46	−12.23	1.85
Mortoniodendron-Pentaplaris	Tilioideae-Malvoideae	0.50	0.50	−34.70	−22.46	−12.23	1.85
Guichenotia-Corchoropsis, Guichenotia-Cheirolaena	Byttnerioideae-Dombeyoideae	0.93	0.06	−79.60	−67.69	−11.91	0.93
Guichenotia-Durio	Byttnerioideae-*Durio*	0.93	0.06	−79.60	−67.69	−11.91	0.93
Guichenotia-Heritiera, Guichenotia-Brachychiton	Byttnerioideae-Sterculioideae	0.93	0.06	−79.60	−67.69	−11.91	0.93
Mortoniodendron-Guichenotia, Guichenotia-Tilia	Byttnerioideae-Tilioideae	0.93	0.06	−79.60	−67.69	−11.91	0.93
Reevesia-Durio	Helicterioideae-*Durio*	0.97	0.03	−211.37	−196.92	−14.45	0.93
Cheirolaena-Tilia, Corchoropsis-Tilia	Dombeyoideae-Tilioideae	0.97	0.03	−207.21	−194.06	−13.15	0.93
Durio-Huberodendron Durio-Pachira	Bombacoideae-*Durio*	0.97	0.03	−187.35	−174.36	−12.99	0.93
Reevesia-Corchoropsis, Reevesia-Cheirolaena	Helicterioideae-Dombeyoideae	0.96	0.04	−160.88	−149.11	−11.77	0.93
Corchoropsis-Mortoniodendron, Mortoniodendron-Cheirolaena	Dombeyoideae-Tilioideae	0.97	0.03	−228.43	−216.54	−11.89	0.93
Mortoniodendron-Reevesia, Reevesia-Tilia	Helicterioideae-Tilioideae	0.97	0.03	−176.27	−163.27	−13.00	0.93
Reevesia-Heritiera, Reevesia-Brachychiton	Helicterioideae-Sterculioideae	0.96	0.04	−166.77	−152.58	−14.19	0.93
Carpodiptera-Reevesia, Reevesia-Brownlowia	Helicterioideae-Brownlowioideae	0.97	0.03	−172.49	−160.95	−11.54	0.93
Huberodendron-Glyphaea	Grewioideae-Bombacoideae	0.87	0.13	−43.04	−30.94	−12.10	0.93
Pentaplaris-Glyphaea, Duboscia-Pentaplaris	Grewioideae-Malvoideae	0.50	0.50	−24.16	−3.71	−20.45	0.93
Huberodendron-Abutilon	Malvoideae-Bombacoideae	0.97	0.03	−187.35	−174.36	−12.99	0.93
Theobroma-Pentaplaris	Byttnerioideae-Malvoideae	0.50	0.50	−24.16	−5.16	−19.00	0.93
Reevesia-Pachira	Helicterioideae-Bombacoideae	0.72	0.28	−246.36	**−256.52**	10.16	0.10
Reevesia-Huberodendron	Helicterioideae-Bombacoideae	0.79	0.21	−204.46	**−214.47**	10.01	0.06
Reevesia-Abutilon	Helicterioideae-Malvoideae	0.84	0.16	−204.18	**−214.43**	10.25	0.05

Summary of the QuIBL results considering only significant (>10 dBIC) values. For simplicity, we refer to the included species by their genus name. ILS proportion reports the proportion of loci with ILS signal, whereas non-ILS proportion indicates the proportion of loci that show additionally an introgression pattern. Bayesian information criterion values are reported for BIC1 and BIC2, where BIC1 is the ILS-only model, while BIC2 model considers ILS and introgression. dBIC is the difference between BIC2 and BIC1 values and it was considered a measure of significance. Total non-ILS proportion is the number of introgressed loci between the two species in the taxa pairs. Bold numbers correspond to models where ILS was preferred over introgression. For detailed results, see **Supplementary Table S4**.

the aim of the research was to generate a phylogenetic hypothesis of Malvaceae that accounts for discordance and heterogeneity among nuclear loci, for which we evaluated the extent and potential sources of discordance, and examined its effect on estimating divergence times.

We sampled all nine subfamilies (40% of all genera) and analyzed the phylogenetic relationships with nuclear data; therefore, this is the first study showing inter- and intra-subfamilial relationships with nuclear sequences. We found high proportions of ILS (**Table 1**), which reduce accuracy in the "single-site" coalescence methods, such as SVDquartets (Chou et al., 2015). Thus, we will base our discussion on further discuss the results from ASTRAL (**Figure 1B**). Either by concatenating all loci or with coalescence, the concordant, highly supported (1 LPP) deep relationships include the two major clades of Malvaceae, Byttneriina, and Malvadendrina. Although there is low degree of introgression between Byttneriina and Malvadendrina members, as our analysis of discordance shows (**Figure 2**) and as was previously found in a reduced nuclear loci sample (Hernández-Gutiérrez et al., 2021), there is strong support of them being two, relatively old, and independent lineages. Other strongly supported relationships pertain to a clade conformed by four subfamilies, Sterculioideae, Tilioideae, and Brownlowioideae + Dombeyoideae (1 LPP); the Malvatheca clade; and within Malvatheca, *Chiranthodendron* as the sister to the remaining species of the clade, and *Ochroma* is subsequently sister to Bombacoideae + Malvoideae (**Figure 1B**).

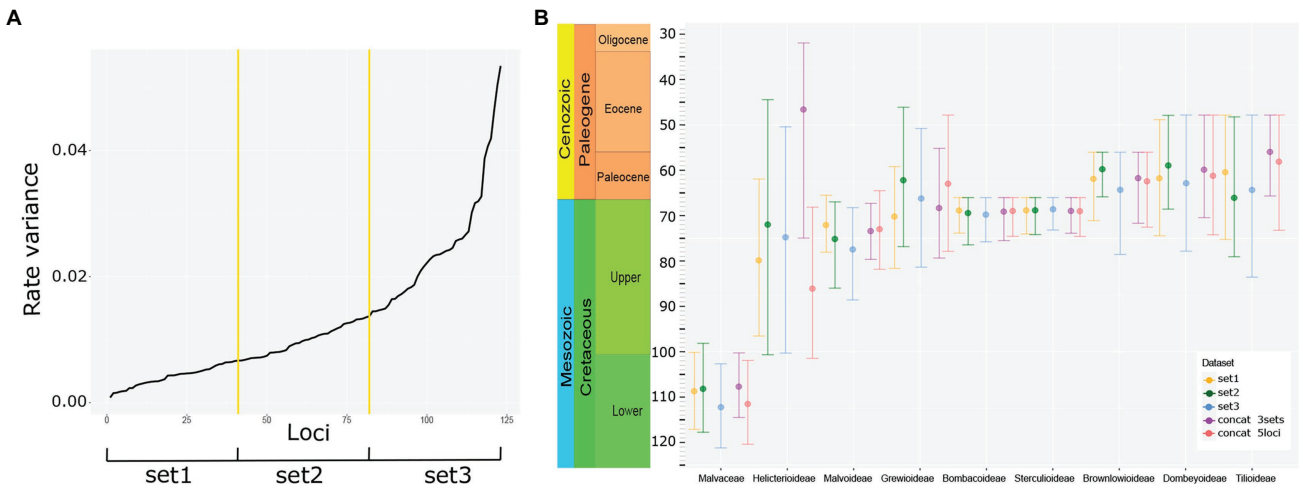


FIGURE 3 | Molecular rate heterogeneity and age estimates. **(A)** Molecular rate variance for 123 loci, where the first tercile has lower variance (set1) and the third tercile has a higher variance (set3) as calculated with SortaDate. **(B)** Crown age estimates derived from five different datasets for Malvaceae and subfamilies, obtained with BEAST2. Circles show mean values, while the bars show 95% highest posterior density (HPD) intervals.

Phylogenetic discordance among nuclear loci is evidence of the possible processes that took place in the history of Malvaceae. Hence, rather than a highly supported and completely resolved topology, here we aimed to obtain an estimate of the extent of phylogenetic discordance in the intricate history of reticulation and rapid diversification characteristic of the family (Conover et al., 2019). Through the coalescence methods, it was possible to locate specific points deep in the phylogeny where discordance is higher: among the four subfamilies Sterculioideae, Tilioideae, Dombeyoideae, and Brownlowioideae; the placement of Helicterioideae; and the Byttnerioideae groups (**Figure 1**). We discuss each of these three cases:

Three previous studies using plastomes have yielded conflicting, highly supported results. For example, in Conover et al. (2019), Dombeyoideae is sister to a clade formed by Sterculioideae + Tilioideae and Malvatheca; in Wang et al. (2020), Sterculioideae is sister only to Tilioideae + Dombeyoideae; and in Cvetković et al. (2021), Sterculioideae is sister to Tilioideae + Dombeyoideae, Brownlowioideae, and Malvatheca. With nuclear data, the study by Hernández-Gutiérrez et al. (2021) shows Sterculioideae as sister to Brownlowioideae, and, similarly to Conover et al. (2019), Dombeyoideae as sister to the remaining Malvadendrina subfamilies, albeit with low support. In the present study, we found these relationships: Sterculioideae + Tilioideae (0.83 LPP), Dombeyoideae + Brownlowioideae (0.99 LPP), these four subfamilies forming a highly supported clade (1 LPP). The low support of Sterculioideae as sister to Tilioideae, and the general conflict of these four subfamilies observed in the previous and the present study is explained by a strong signal of ILS present between each of these four subfamilies and other subfamilies, for example, Malvoideae and Byttnerioideae (**Table 1**) and combined with a relatively little proportion of reticulation with a member of Malvatheca experienced early in their diversification (**Figure 2**; **Table 1**).

In most of the phylogenetic analyses, we retrieved Helicterioideae (excluding *Durio*) as the sister group of the remaining Malvadendrina (0.61 LPP), which is congruent with plastome phylogenetic analyses (Conover et al., 2019; Cvetković et al., 2021; Wang et al., 2021). Our results yield a significantly preferred model of ILS when Helicterioideae is associated to Malvatheca (**Table 1**); thus, the discordant placement of Helicterioideae is probably caused by ILS only (**Figure 2**). In all our analyses, *Durio zibethinus* appears outside Helicterioideae and is sister to the clade comprising Sterculioideae, Tilioideae, Dombeyoideae, and Brownlowioideae. This placement is possibly derived from reticulation with members of other subfamilies given that phylogenetic discordance analyses show a high degree of introgression between *Durio* and Bombacoideae and *Durio* and Byttnerioideae (**Table 1**). The former introgression event had been previously inferred (Conover et al., 2019), all of which might be causing that nuclear information leads to such phylogenetic results. This needs to be explicitly examined with a denser sampling of species from the genus *Durio* and tribe Durioneae.

In this study, Byttnerioideae appeared as paraphyletic (**Figure 1**), except in the concatenated analysis where the subfamily was monophyletic (72 BS; **Supplementary Figure S1**). However, previous analyses with few plastid molecular markers, but a well-represented taxon sampling, showed that Byttnerioideae was strongly to moderately supported as a monophyletic group (Whitlock et al., 2001; Richardson et al., 2015; Hernández-Gutiérrez and Magallón, 2019). A group including *Byttneria* and *Leptonichia* (and other species of tribe Byttneriae) are separated from the rest of Byttnerioideae and are more closely related to Grewioideae (**Figure 1**), although with low support (0.27 LPP) and deriving from less than 30 QT (**Figure 1B**). Additional to ILS, the source of discordance in this case seems to derive from a low proportion of introgression between *Guichenotia* and the

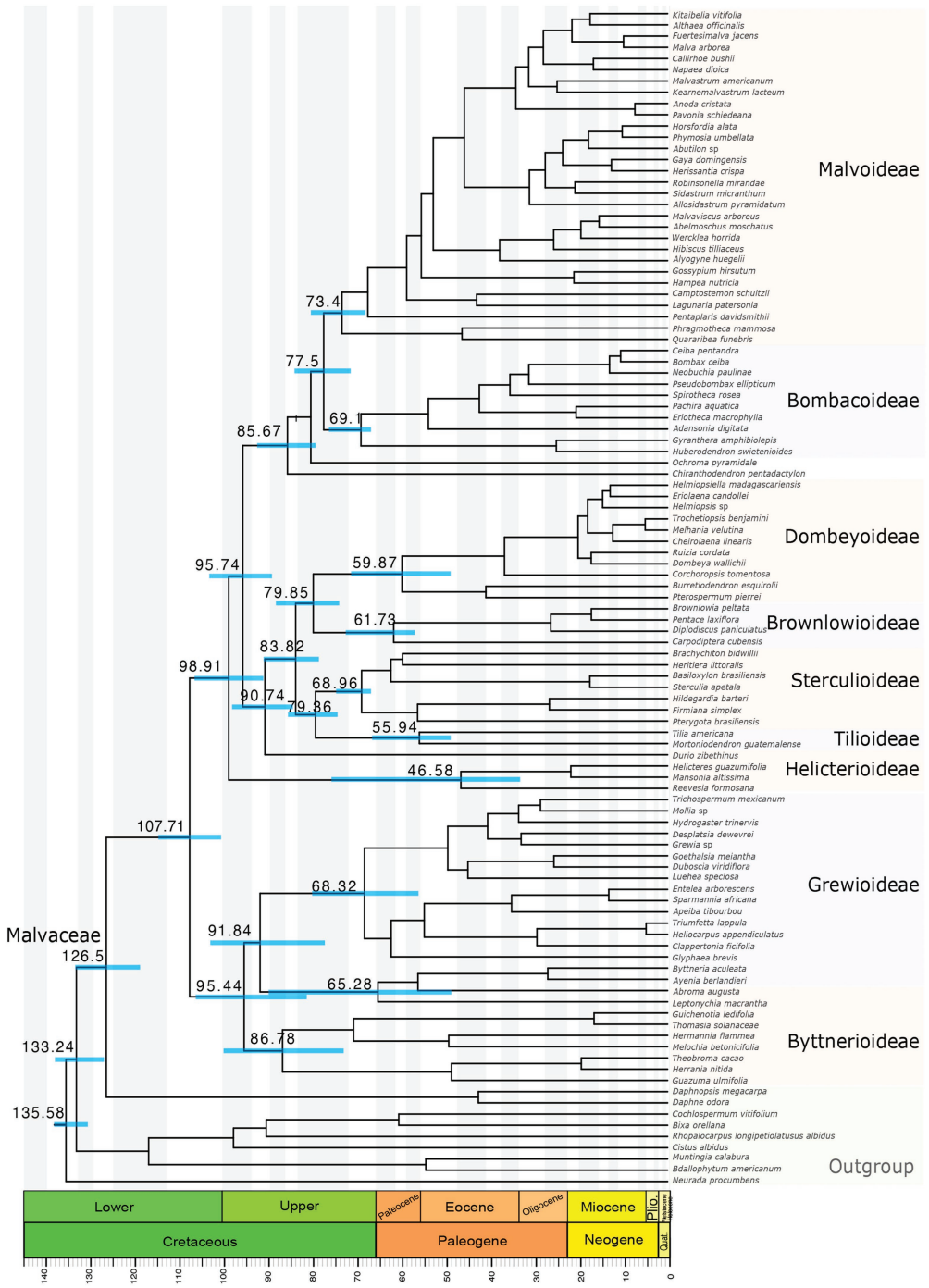


FIGURE 4 | Maximum clade credibility tree derived from the concatenated dataset (123 loci) partitioned by set (set1, set2, and set3). Bars associated to age values are the 95% highest posterior density (HPD) intervals.

common ancestor of Sterculioideae, Tilioideae, Dombeyoideae, and Brownlowioideae (**Figure 2**; **Table 1**) and between *Theobroma*/*Guichenotia* and Malvoideae. This is an area for further research, as plastome analyses have included maximum two genera of this group (Conover et al., 2019; Cvetković et al., 2021; Wang et al., 2021).

Overall, our results indicate that ILS is the main source of phylogenetic discordance in the relationships among subfamilies, but only if combined with different degrees of introgression (**Table 1**). Thus, together these two processes explain the contentious relationships of the major lineages of Malvaceae (**Figure 2**). Analyzing whole-genome multiplications,

Conover et al. (2019) formulated two alternative hypotheses. One considers an allopolyploidization event between dombeyoid and Malvatheca ancestors that gave rise to *Durio*. This hypothesis is somewhat consistent with our findings, but we detected a significant signal of introgression between *Durio* and Bombacoideae (not Malvoideae) and *Reevesia* (Helicterioideae). A potential allopolyploidization between the ancestors of Helicterioideae and Malvatheca/Bombacoideae may have caused the position of *Durio* apart from the rest of Helicterioideae found with our nuclear loci. The second hypothesis considers that Malvatheca originated *via* allopolyploidization between Sterculioideae + Tilioideae and Helicterioideae. This scenario is supported by a consistent introgression signal between Malvoideae and Sterculioideae + Tilioideae, but introgression was also detected between the four subfamilies Sterculioideae + Tilioideae + Dombeyoideae + Brownlowioideae and Bombacoideae. Therefore, it is possible that the observed signal in the genomes comes from a reticulation event involving the ancestors of the four subfamilies and the ancestor of Malvatheca. Moreover, considering Byttnerioideae and Grewioideae adds to the formulated hypotheses by Conover et al. (2019) a more complicated component, which is introgression between Malvadendrina and Byttneriina members obscured by a generalized ILS (**Figure 2**). How these past events shaped the morphological evolution of Malvaceae is now an interesting question to address, since it has been proved that floral traits acquired by introgression in baobabs might have led to adaptive evolution (Karimi et al., 2020).

A potential limitation of our analyses on the source of phylogenetic discordance relies on the assumption that gene trees are correctly estimated, because they were used to estimate the ASTRAL species tree that subsequently was used to compare the discordant topologies when interpreting QuIBL results. One particular aspect of gene tree inference concerns the collapsing of low supported bipartitions, where it has been identified that different collapsing methods have severe impacts on tree reconstruction (Simmons and Gatesy, 2021). Furthermore, as we discuss further in the next section, molecular rate heterogeneity has a strong impact on phylogenetic inferences in general, with new evidence on its impact on species tree estimation (Vankan et al., 2021). In the present study, we considered the heterogeneity in rates for divergence times, but not for the species tree estimation.

Molecular Rate Heterogeneity and Discordance: Implications for Molecular Dating

Rate heterogeneity in Malvaceae was previously quantified within Malvatheca (Baum et al., 2004) and among the genomes of cotton, durian, and cacao (Wang et al., 2019a), where shifts in molecular evolutionary rate were many unit fold between Malvoideae and Bombacoideae (Baum et al., 2004) and between cotton and either durian or cacao (Wang et al., 2019a). It was thus expected to find high heterogeneity in our nuclear loci sampling (**Figure 3A**). Molecular rate heterogeneity is a long-recognized factor influencing both phylogenetic inference and divergence time estimation (Yang, 1995; Sanderson, 1997; Thorne et al., 1998), an influence that is exacerbated using hundreds of loci (Smith et al., 2018; Dornburg et al., 2019). Particularly important is the selection of loci and the assumptions on molecular clock models (Carruthers et al., 2020). Here, we used a "gene-shopping" approach to categorize loci by their rate, and then form three sets, of low, moderate, and high heterogeneity, each with 41 concatenated loci (**Figure 3A**). We also concatenated all loci into a single alignment partitioned by set. A final alignment was considered using the five loci with the lowest rate variance, which represents a conservative analysis given its homogeneity in molecular rate. Divergence time estimates show similar results among the five analyses, but the few observed substantial differences are probably due to phylogenetic discordances. Moreover, older ages were obtained from the third tercile of rate variance (set3 in **Figures 3A,B**), demonstrating that, although close and overlapping results, rate variance influences the general pattern of divergence times.

We found congruent age estimates possibly due to the number of calibrations we used, as it is known that when the heterogeneity in rate estimates is large multiple calibrations may constrain the estimates (Ho and Phillips, 2009). Divergence times here obtained are older than in Hernández-Gutiérrez and Magallón (2019), except for Tilioideae, which is younger. The difference might be related to numerous factors, such as molecular rate, taxon sampling, and phylogenetic relationships, but possibly mostly because the secondary calibration here applied to the Malvales, which was derived from the Ramírez-Barahona et al. (2020) study, is older than the one applied in the previous analysis. Wang et al. (2021) performed a divergence time estimation of Malvaceae and its subfamilies showing younger ages, probably due to the young secondary calibration, which was based on an analysis with a secondary calibration and Pure birth (Yule, 1924) tree diversification model (Richardson et al., 2015). Surprisingly, the here estimated crown age of Malvaceae roughly coincides with that estimated in Cvetković et al. (2021), but subfamilial ages in the present study are older possibly due to the larger taxon sampling.

In this study, we found that nuclear loci are highly variable in molecular rate and in phylogenetic histories, translated in high heterogeneity, and phylogenetic discordance, in particular, during the early diversification of the subfamilies. However, we were able to detect that ILS and different extents of introgression underlie this discordance and that rate heterogeneity slightly affects divergence time estimation due possibly to the combined information from the calibration priors. We also found that Helicterioideae and Byttnerioideae need to be further sampled and analyzed in the context to the remaining Malvadendrina groups and the relationships within them.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the NCBI Sequence Read Archive repository in the BioProject accession PRJNA815625.

AUTHOR CONTRIBUTIONS

RH-G, CB, CGM, MPC, EFM, and SM: field collection. RH-G, CB, CGM, and EL: laboratory procedure. AL: development and application of bioinformatic pipelines for raw data processing. RH-G: data analyses and visualization of results and writing of the first draft. All authors reviewed and edited the draft and agreed to the submitted version of the manuscript.

FUNDING

We gratefully acknowledge funding provided by Programa de Apoyos a Proyectos de Investigación e Innovación Tecnológica of the Universidad Nacional Autónoma de México (UNAM) PAPIIT IG200316 and Fronteras de la Ciencia, Consejo Nacional de Ciencia y Tecnología (CONACyT) project number 2016-01-1867, both granted to SM. RH-G received a doctoral scholarship from CONACyT (407103/288658) and received the Elizabeth E. Bascom Fellowship for Latin American Women from the Missouri Botanical Garden, the American Society of Plant Taxonomists (ASPT) Graduate Research Grant, and the International Association for Plant Taxonomy (IAPT) Research Grant.

ACKNOWLEDGMENTS

We are very grateful to the Instituto Nacional de Biodiversidad del Ecuador (INABIO) for logistic support during herbarium and field work, and to Ministerio del Ambiente, Agua y Transición Ecológica del Ecuador (MAATE) for granting a permit for scientific collection of samples (permit number: MAE-DNB-CM-2016-0045). We thank Rafael Torres, Álvaro Campos, Arturo de Nova, and Luna Sánchez for their help in fieldwork in Mexico; Alex Popovkin for his great support in fieldwork in Brazil; Ricardo Perdiz for sharing collecting localities in Brazil; Adriana Benítez for her support in fieldwork and in bioinformatic work; Gerardo Salazar, Itzi Fragoso and Mario Ishiki for sampling some species in field; Miriam Miyagi for helping with QuIBL; Lidia Cabrera (Laboratorio de Biología Molecular, LANABIO IB-UNAM) for supporting in laboratory work; and Alfredo Wong (IB-UNAM) for support with BEAGLE server. We are grateful to the institutions and people that permitted and helped in the sampling from living and herbarium collections: Missouri Botanical Garden (Jim Solomon, Rebecca Sucher), Royal Botanic Gardens Kew (Sara Edwards, Bente Klitgaard, Felix Forest, Olivier Murin, Alan Paton), Royal Botanic Garden Edinburgh (Toby Pennington, Peter Brownless), Botanic Garden Meise (Frank Van Caekenberghe), and the National Herbarium of Mexico (MEXU; María del Rosario García, Verónica Juárez, Laura Calvillo, Angélica Ramírez, Gilda Ortiz, Alberto Reyes, David Gernandt). We acknowledge the reviewers for their comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.850521/full#supplementary-material

REFERENCES

Alverson, W. S., Karol, K. G., Baum, D. A., Chase, M. W., Swensen, S. M., McCourt, R., et al. (1998). Circumscription of the Malvales and relationships to other Rosidae: evidence from rbcL sequence data. *Am. J. Bot.* 85, 876–887.

Alverson, W. S., Whitlock, B. A., Nyffeler, R., Bayer, C., and Baum, D. A. (1999). Phylogeny of the core Malvales: evidence from *ndhF* sequence data. *Am. J. Bot.* 86, 1474–1486. doi: 10.2307/2656928

Angelis, K., Álvarez-Carretero, S., Dos Reis, M., and Yang, Z. (2018). An evaluation of different partitioning strategies for Bayesian estimation of species divergence times. *Syst. Biol.* 67, 61–77. doi: 10.1093/sysbio/syx061

Baum, D., Smith, S., Yen, A., Alverson, W., Nyffeler, R., Whitlock, B., et al. (2004). Phylogenetic relationships of Malvatheca (Bombacoideae and Malvoideae; Malvaceae sensu lato) as inferred from plastid DNA sequences. *Am. J. Bot.* 91, 1863–1871. doi: 10.3732/ajb.91.11.1863

Bayer, C., Fay, M. F., De Bruijn, A. Y., Savolainen, V., Morton, C. M., Kubitzki, K., et al. (1999). Support for an expanded family concept of Malvaceae within a recircumscribed order Malvales: a combined analysis of plastid atpB and rbcL DNA sequences. *Bot. J. Linn. Soc.* 129, 267–303. doi: 10.1111/j.1095-8339.1999.tb00505.x

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Brown, J. W., Walker, J. F., and Smith, S. A. (2017). Phyx: phylogenetic tools for unix. *Bioinformatics* 33, 1886–1888. doi: 10.1093/bioinformatics/btx063

Buddenhagen, C., Lemmon, A. R., Lemmon, E. M., Bruhl, J., Cappa, J., Clement, W. L., et al. (2016). Anchored phylogenetics of angiosperms I: assessing the robustness of phylogenetic estimates. bioRxiv 086298. doi:10.1101/086298.

Cai, L., Xi, Z., Moriarty Lemmon, E., Lemmon, A. R., Mast, A., Buddenhagen, E., et al. (2021). The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales. *Syst. Biol.* 70, 491–507. doi: 10.1093/sysbio/syaa083

Carruthers, T., Sanderson, M. J., and Scotrland, R. W. (2020). The implications of lineage-specific rates for divergence time estimation. *Syst. Biol.* 69, 660–670. doi: 10.1093/sysbio/syz080

Chifman, J., and Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30, 3317–3324. doi: 10.1093/bioinformatics/btu530

Chou, J., Gupta, A., Yaduvanshi, S., Davidson, R., Nute, M., Mirarab, S., et al. (2015). A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics* 16:S2. doi: 10.1186/1471-2164-16-S10-S2

Conover, J. L., Karimi, N., Stenz, N., Ané, C., Grover, C. E., Skema, C., et al. (2019). A Malvaceae mystery: a mallow maelstrom of genome multiplications and maybe misleading methods? *J. Integr. Plant Biol.* 61, 12–31. doi: 10.1111/jipb.12746

Costa, L., Oliveira, A., Carvalho-Sobrinho, J., and Souza, G. (2017). Comparative cytomolecular analyses reveal karyotype variability related to biogeographic and species richness patterns in Bombacoideae (Malvaceae). *Plant Syst. Evol.* 303, 1131–1144. doi: 10.1007/s00606-017-1427-6

Cvetković, T., Areces-Berazain, F., Hinsinger, D. D., Thomas, D. C., Wieringa, J. J., Ganesan, S. K., et al. (2021). Phylogenomics resolves deep subfamilial relationships in Malvaceae s.l. *G3* 11:jkab136. doi: 10.1093/g3journal/jkab136

Dornburg, A., Su, Z., and Townsend, J. P. (2019). Optimal rates for phylogenetic inference and experimental design in the era of genome-scale data sets. *Syst. Biol.* 68, 145–156. doi: 10.1093/sysbio/syy047

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88. doi: 10.1371/journal.pbio.0040088

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., et al. (2019). Genomic architecture and introgression shape a butterfly radiation. *Science* 366, 594–599. doi: 10.1126/science.aaw2090

Fragoso-Martínez, I., Salazar, G. A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Moriarty Lemmon, E., et al. (2017). A pilot study applying the plant anchored hybrid enrichment method to New World sages (*Salvia* subgenus Calosphace; Lamiaceae). 25th Anniv. Issue. *Mol. Phylogenet. Evol.* 117, 124–134. doi: 10.1016/j.ympev.2017.02.006

Granados Mendoza, C., Jost, M., Hágsater, E., Magallón, S., van den Berg, C., Moriarty Lemmon, E., et al. (2020). Target nuclear and off-target plastid hybrid enrichment data inform a range of evolutionary depths in the orchid genus Epidendrum. *Front. Plant Sci.* 10:1761. doi: 10.3389/fpls.2019.01761

Hamilton, C. A., Lemmon, A. R., Moriarty Lemmon, E., and Bond, J. E. (2016). Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. *BMC Evol. Biol.* 16:212. doi: 10.1186/s12862-016-0769-y

Heibl, C., Cusimano, N., and Krah, F.-S. (2019). Package 'ips'. Interfaces to Phylogenetic Software in R.

Hernández-Gutiérrez, R., Granados Mendoza, C., and Magallón, S. (2021). Low-copy nuclear genes reveal new evidence of incongruence in relationships within Malvaceae s.l. *Syst. Bot.* 46, 1042–1052. doi: 10.1600/036364421X16370109698551

Hernández-Gutiérrez, R., and Magallón, S. (2019). The timing of Malvales evolution: incorporating its extensive fossil record to inform about lineage diversification. *Mol. Phylogenet. Evol.* 140:106606. doi: 10.1016/j.ympev.2019.106606

Ho, S. Y. W., and Phillips, M. J. (2009). Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst. Biol.* 58, 367–380. doi: 10.1093/sysbio/syp035

Jost, M., Samain, M.-S., Marques, I., Graham, S. W., and Wanke, S. (2021). Discordant phylogenomic placement of Hydnoraceae and Lactoridaceae within Piperales using data from all three genomes. *Front. Plant Sci.* 12:642598. doi: 10.3389/fpls.2021.642598

Junier, T., and Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX Shell. *Bioinformatics* 26, 1669–1670. doi: 10.1093/bioinformatics/btq243

Karimi, N., Grover, C. E., Gallagher, J. P., Wenderl, J. F., Ané, C., and Baum, D. A. (2020). Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (Adansonia; Bombacoideae; Malvaceae). *Syst. Biol.* 69, 462–478. doi: 10.1093/sysbio/syz073

Katoh, K., and Standley, D. M. (2013). MAFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Koenen, E. J. M., Ojeda, D. I., Steeves, R., Migliore, J., Bakker, F. T., Wieringa, J. J., et al. (2020). Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. *New Phytol.* 225, 1355–1369. doi: 10.1111/nph.16290

Lanfear, R., Calcott, B., Ho, S. Y. W., and Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29, 1695–1701. doi: 10.1093/molbev/mss020

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2016). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* 34, 772–msw773. doi: 10.1093/molbev/msw260

Lee, E. K., Cibrian-Jaramillo, A., Kolokotronis, S.-O., Katari, M. S., Stamatakis, A., Ott, M., et al. (2011). A functional phylogenomic view of the seed plants. *PLoS Genet.* 7:e1002411. doi: 10.1371/journal.pgen.1002411

Lemmon, A. R. (2015). U.S. Patent Application No. 14/524,614.

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* 2010:pdb.prot5448. doi: 10.1101/pdb.prot5448

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the cipres science gateway for inference of large phylogenetic trees." in *Proceedings of the Gateway Computing Environments Workshop (GCE)*, November 14, 2010; New Orleans, LA, 1–8.

Morales-Briones, D. F., Kadereit, G., Tefarikis, D. T., Moore, M. J., Smith, S. A., Brockington, S. F., et al. (2021). Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae s.l. *Syst. Biol.* 70, 219–235. doi: 10.1093/sysbio/syaa066

Nie, Y., Foster, C. S. P., Zhu, T., Yao, R., Duchene, D. A., Ho, S. Y. W., et al. (2020). Accounting for uncertainty in the evolutionary timescale of green plants through clock-partitioning and fossil calibration strategies. *Syst. Biol.* 69, 1–16. doi: 10.1093/sysbio/syz032

Nyffeler, R., Bayer, C., Alverson, W. S., Yen, A., Whitlock, B. A., Chase, M. K., et al. (2005). Phylogenetic analysis of the malvadendrina clade (Malvaceae s.l.) based on plastid DNA sequences. *Org. Divers. Evol.* 5, 109–123. doi: 10.1016/j.ode.2004.08.001

Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., et al. (2012). Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 492, 423–427. doi: 10.1038/nature11798

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Moriarty Lemmon, E., et al. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526, 569–573. doi: 10.1038/nature15697

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ (Accessed December, 2020).

Ramírez-Barahona, S., Sauquet, H., and Magallón, S. (2020). The delayed and geographically heterogeneous diversification of flowering plant families. *Nat. Ecol. Evol.* 4, 1232–1238. doi: 10.1038/s41559-020-1241-3

Richardson, J. E., Whitlock, B. A., Meerow, A. W., and Madriñán, S. (2015). The age of chocolate: a diversification history of Theobroma and Malvaceae. *Front. Ecol. Evol.* 3:120. doi: 10.3389/fevo.2015.00120

Rokyta, D. R., Lemmon, A. R., Margres, M. J., and Aronow, K. (2012). The venom-gland transcriptome of the eastern diamondback rattlesnake (Crotalus adamanteus). *BMC Genomics* 13, 312–323. doi: 10.1186/1471-2164-13-312

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. PMID: 3447015

Sanderson, M. J. (1997). A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* 14, 1218–1231. doi: 10.1093/oxfordjournals.molbev.a025731

Simmons, M. P., and Gatesy, J. (2021). Collapsing dubiously resolved gene-tree branches in phylogenomic coalescent analyses. *Mol. Phylogenet. Evol.* 158:107092. doi: 10.1016/j.ympev.2021.107092

Smith, S. A., Brown, J. W., and Walker, J. F. (2018). So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLoS One* 13:e0197433. doi: 10.1371/journal.pone.0197433

Sorenson, M. D., and Franzosa, E. A. (2007). *TreeRot, Version 3*. Boston, MA: Boston University.

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

Swofford, D. L. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sunderland, Mass: Sinauer Associates.

The Plant List (2021). Pollinator demonstration garden at pinewood lake park. Available at: http://www.theplantlis.org/ (Accessed September, 2021).

Thorne, J. L., Kishino, H., and Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15, 1647–1657. doi: 10.1093/oxfordjournals.molbev.a025892

Uyeda, J., and Harmon, L. (2020). treeplyr: 'dplyr' functionality for matched tree and data objects. Available at: https://github.com/uyedaj/treeplyr (Accessed November, 2020).

Vankan, M., Ho, S. Y. W., and Duchêne, D. A. (2021). Evolutionary rate variation among lineages in gene trees has a negative impact on species-tree inference. *Syst. Biol.* 71, 490–500. doi: 10.1093/sysbio/syab051

Wang, L.-G., Lam, T. T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., et al. (2020). Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* 37, 599–603. doi: 10.1093/molbev/msz240

Wang, J.-H., Moore, M. J., Wang, H., Zhu, Z.-X., and Wang, H.-F. (2021). Plastome evolution and phylogenetic relationships among Malvaceae subfamilies. *Gene* 765:145103. doi: 10.1016/j.gene.2020.145103

Wang, N., Yang, Y., Moore, M. J., Brockington, S. F., Walker, J. F., Brown, J. W., et al. (2019b). Evolution of Portulacineae marked by gene tree conflict and gene family expansion associated with adaptation to harsh environments. *Mol. Biol. Evol.* 36, 112–126. doi: 10.1093/molbev/msy200

Wang, J., Yuan, J., Yu, J., Meng, F., Sun, P., Li, Y., et al. (2019a). Recursive paleohexaploidization shaped the durian genome. *Plant Physiol.* 179, 209–219. doi: 10.1104/pp.18.00921

Wanke, S., Granados Mendoza, C., Müller, S., Paizanni Guillén, A., Neinhuis, C., Lemmon, A. R., et al. (2017). Recalcitrant deep and shallow nodes in Aristolochia (Aristolochiaceae) illuminated using anchored hybrid enrichment. 25th Anniv. Issue. *Mol. Phylogenet. Evol.* 117, 111–123. doi: 10.1016/j.ympev.2017.05.014

Whitlock, B. A., Bayer, C., and Baum, D. A. (2001). Phylogenetic relationships and floral evolution of the Byttnerioideae ("Sterculiaceae" or Malvaceae s.l.) based on sequences of the chloroplast gene ndhF. *Syst. Bot.* 26, 420–437. doi: 10.1043/0363-6445-26.2.420

Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics* 139, 993–1005. doi: 10.1093/genetics/139.2.993

Yule, G. U. (1924). A mathematical theory of evolution: based on the conclusions of Dr. J.C. Willis. *Philos. Trans. R. Soc. B* 213, 21–87. doi: 10.1098/rstb.1925.0002

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19(S6):153. doi: 10.1186/s12859-018-2129-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 27 May 2022
doi: 10.3389/fpls.2022.869583

Pervasive Phylogenomic Incongruence Underlies Evolutionary Relationships in Eyebrights (*Euphrasia*, Orobanchaceae)

Phen Garrett[1], Hannes Becher[2], Galina Gussarova[3,4,5], Claude W. dePamphilis[6], Rob W. Ness[7], Shyam Gopalakrishnan[1] and Alex D. Twyford[2,8]**

[1]GLOBE Institute, University of Copenhagen, Copenhagen, Denmark, [2]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, [3]Natural History Museum, University of Oslo, Oslo, Norway, [4]Botany Department, Faculty of Biology and Soil Science, St Petersburg State University, St Petersburg, Russia, [5]Tromsø University Museum, University of Tromsø, Tromsø, Norway, [6]Department of Biology and Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, United States, [7]Department of Biology, University of Toronto Mississauga, Mississauga, ON, Canada, [8]Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom

OPEN ACCESS

Edited by:
Susann Wicke, Humboldt University of Berlin, Germany

Reviewed by:
Christoph Oberprieler, University of Regensburg, Germany
Richard Hodel, Smithsonian National Museum of Natural History (SI), United States
Wen-Bin Yu, Xishuangbanna Tropical Botanical Garden (CAS), China

***Correspondence:**
Phen Garrett
phengarrett@gmail.com
Alex D. Twyford
alex.twyford@ed.ac.uk

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *04 February 2022*
Accepted: *19 April 2022*
Published: *27 May 2022*

Citation:
Garrett P, Becher H, Gussarova G, dePamphilis CW, Ness RW, Gopalakrishnan S and Twyford AD (2022) Pervasive Phylogenomic Incongruence Underlies Evolutionary Relationships in Eyebrights (Euphrasia, Orobanchaceae). Front. Plant Sci. 13:869583. doi: 10.3389/fpls.2022.869583

Disentangling the phylogenetic relationships of taxonomically complex plant groups is often mired by challenges associated with recent speciation, hybridization, complex mating systems, and polyploidy. Here, we perform a phylogenomic analysis of eyebrights (*Euphrasia*), a group renowned for taxonomic complexity, with the aim of documenting the extent of phylogenetic discordance at both deep and at shallow phylogenetic scales. We generate whole-genome sequencing data and integrate this with prior genomic data to perform a comprehensive analysis of nuclear genomic, nuclear ribosomal (nrDNA), and complete plastid genomes from 57 individuals representing 36 *Euphrasia* species. The species tree analysis of 3,454 conserved nuclear scaffolds (46 Mb) reveals that at shallow phylogenetic scales postglacial colonization of North Western Europe occurred in multiple waves from discrete source populations, with most species not being monophyletic, and instead combining genomic variants from across clades. At a deeper phylogenetic scale, the *Euphrasia* phylogeny is structured by geography and ploidy, and partially by taxonomy. Comparative analyses show Southern Hemisphere tetraploids include a distinct subgenome indicative of independent polyploidy events from Northern Hemisphere taxa. In contrast to the nuclear genome analyses, the plastid genome phylogeny reveals limited geographic structure, while the nrDNA phylogeny is informative of some geographic and taxonomic affinities but more thorough phylogenetic inference is impeded by the retention of ancestral polymorphisms in the polyploids. Overall our results reveal extensive phylogenetic discordance at both deeper and shallower nodes, with broad-scale geographic structure of genomic variation but a lack of definitive taxonomic signal. This suggests that *Euphrasia* species either have polytopic origins or are maintained by narrow genomic regions in the face of extensive homogenizing gene flow. Moreover, these results suggest genome skimming will not be an effective extended barcode to identify species in groups such as *Euphrasia*, or many other postglacial species groups.

Keywords: phylogeny, discordance, *Euphrasia*, taxonomic complexity, plastid, whole-genome sequencing

INTRODUCTION

Gene tree discordance is a pervasive feature of plant phylogenies, with numerous studies revealing diverse and conflicting topologies among loci within a genome (Stull et al., 2020; Rose et al., 2021; Wagner et al., 2021). While discordance is frequently seen as a barrier to species tree reconstruction and an impediment to taxonomic and systematic research, characterizing discordance can provide major insights into evolutionary processes. For example, discordance at deep phylogenetic scales can be indicative of hybridization that has promoted major species radiations (Marques et al., 2019), while discordance at shallow phylogenetic scales can reveal contemporary population processes, such as the balance between drift, gene flow, and selection in the maintenance of genetic variation (Lee-Yaw et al., 2019). Phylogenetic discordance is likely to be most prevalent in certain plant groups, particularly those characterized by hybridization, polyploidy, and/or recent (often postglacial) speciation (Squirrell et al., 2002; Wagner et al., 2021). This includes taxonomically complex plant groups (*sensu* Ennos et al., 2005), such as sedges (*Carex*), willows (*Salix*), and *Epipactis* orchids, where discrete species are often hard to define.

Phylogenomic studies of taxonomically complex groups are fraught with difficulties, with these poorly studied groups often lacking genomic resources, such as reference genomes, and with bioinformatic issues associated with the analysis of polyploids (Brandrud et al., 2020). Despite these challenges, the emergence of low-cost genomic sequencing, coupled with bioinformatic tools for the analysis of large and complex phylogenomic data sets, makes these issues ever more tractable. For example, genomic data can be generated from expertly determined herbarium samples even if the DNA shows evidence of degradation (Bakker et al., 2016), providing a phylogenetic context and taxonomic framework for interpreting relationships in taxonomically complex groups. Moreover, even low-coverage genomic data, such as genome skimming (Straub et al., 2012), can be useful for recovering multiple independent subcellular genomes, such as the plastid, mitochondrial, and nuclear ribosomal DNA (nrDNA). Studying haploid organellar genomes, as well as nrDNA where repeats are generally expected to be homogenized within an individual *via* concerted evolution (Xu et al., 2017), circumvents many issues associated with polyploid phylogenetics and provides an opportunity to compare phylogenetic signal between genomes with conflicting modes of inheritance (Rieseberg and Soltis, 1991). If nuclear genomic data can also be recovered this may facilitate more detailed characterization of a groups' evolutionary history, such as investigating the evolutionary history of the two or more composite "subgenomes" in allopolyploids (Chen et al., 2020). Despite this promise, few studies to date have investigated phylogenetic discordance in taxonomically complex plant groups (though see Brandrud et al., 2020).

The genus *Euphrasia*, commonly known as eyebrights, are a diverse group comprising 273 annual and perennial species, with a bipolar distribution. Across the genus, there are various ploidy levels, from diploids to dodecaploids, with multiple independent polyploidy events from a base chromosome number of 11 (Gussarova et al., 2008). All species are hemiparasites that attach to, and feed from, a broad range of plant hosts (Yeo, 1964; Brown et al., 2021). The genus is perhaps most renowned for its taxonomic complexity, particularly in Europe (French et al., 2008). The postglacial radiation in northern Europe includes numerous closely related taxa that are extremely challenging to separate based on morphology (Yeo, 1978) or with DNA barcoding (Wang et al., 2018). The small stature of these plants (frequently <10 cm tall) and their phenotypic plasticity (Karlsson, 1984; Zopfi, 1997; Brown et al., 2020), coupled with many traits demonstrating population-level rather than species-level differences due to limited gene flow as a consequence of their selfing or partly selfing mating system (French et al., 2005), all further confound species identification. Furthermore, species show extensive interfertility and a large array of natural hybrids have been recorded in the wild (Stace et al., 2015). The most extensive taxonomic issues have been noted from tetraploid species, though issues are present at all ploidy levels.

Previous studies of the genus have successfully confirmed the monophyly of *Euphrasia* and resolved some broad-scale relationships, though have also faced significant challenges. In terms of broad-scale studies, the largest global phylogeny to date used three plastid regions and the internal transcribed spacer (ITS) of nrDNA generated for 51 species (Gussarova et al., 2008). This study recovered phylogenetic relationships relating to broad-geographic regions and by ploidy. However, for both data sets, there were issues with unresolved species-level relationships, with these particularly pronounced for European taxa in the plastid phylogeny, which were largely unresolved. Moreover, the underlying evolutionary processes shaping the topology were hard to infer with few gene regions, and it may be that nrDNA homogenization or loss of ancestral plastid variants could cause these global phylogenies to deviate from the expected nuclear species tree. At a smaller geographic scale, population genomic sequencing of 18 samples of British *Euphrasia*, with a particular sampling focus of co-occurring species on the small Scottish island of Fair Isle, found that species share extremely similar plastid DNA sequences (>99.8% similarity based on whole plastid genomes), with phylogenetic relationships not closely tracking species boundaries and only weakly clustering by geography (Becher et al., 2020). Diploids and tetraploids were characterized by highly divergent nrDNA arrays (10.8% divergence in ITS sequences), though species-level relationships remained unclear. Here, neither plastid DNA nor nrDNA closely followed the pattern observed across the nuclear genome in these samples.

In this study, we use genomic data to investigate phylogenetic discordance in taxonomically complex *Euphrasia*. To do this we adopt a two-stage strategy. First, we study the enigmatic relationships of postglacial northern European *Euphrasia* species, particularly those present in Britain. We build on a number of previous genetic studies (French et al., 2008; Gussarova et al., 2008; Wang et al., 2018; Becher et al., 2020), generating new sequence data and re-analyzing previous sequences. This microevolutionary focus, aimed at using dense species sampling and the use of multiple individuals per species, allows us to

investigate: (1) phylogenetic relationships and the evidence for recurrent colonization of the British Isles from continental Europe, (2) the nature of species differences and whether hybridizing British *Euphrasia* species are monophyletic. Secondly, we study a sparser sample of diverse species from across the *Euphrasia* phylogeny, with the aim of investigating broader scale macroevolutionary processes. In particular, we look test: (3) whether there is evidence of phylogenetic discordance deep in the *Euphrasia* phylogeny, (4) whether discordance may be a consequence of more complex genome evolutionary dynamics in newly sequenced polyploids. Our approach involves diverse herbarium material used for genomic sequencing, and comparative genomics, to document phylogenetic discordance between independent subcellular genomes and genomic regions (the nuclear genome, nrDNA arrays, plastid genomes).

MATERIALS AND METHODS

Plant Material, DNA Extraction, and Genomic Sequencing

Our phylogenomic analyses included a total of 58 samples, 56 samples from 36 *Euphrasia* species, and two outgroup species, *Bartsia alpina* L. and *Neobartsia chilensis* Uribe-Convers & Tank. This material included a combination of newly sequenced samples and previously generated data, with full sample information provided in **Supplementary Table S1**. To investigate global evolutionary relationships, broad-scale phylogenetic conflict, and diversity in modes of ploidy across the genus, we selected a shallow sample of taxa that maximized representation of geographic regions (including Northern and Southern Hemisphere taxa), and to capture taxonomic diversity and anticipated evolutionary divergence times. To investigate phylogenetic relationships and species cohesion in postglacial northern European *Euphrasia*, particularly in Britain, we used available sequences from a range of different studies to maximize species coverage, and where possible to include multiple individuals per species.

For the broad-scale analysis, we sequenced herbarium material from 17 *Euphrasia* species. Herbarium samples were obtained from the University of Copenhagen (C), the Royal Botanic Garden Edinburgh (E), and Oslo University Herbarium (O). The herbarium samples spanned 1861–2019 and included a broad range of collection localities covering 15 countries including Canada, New Zealand, and Sweden (**Supplementary Table S1**). DNA was extracted from samples using the Qiagen DNEasy Plant Extraction kit. Extractions were quantified using the Qubit 2.0 Fluorometer (Applied Biosystems).

Library building was performed using Copenhagen University's EvoGenomics' in-house BEST protocol (Carøe et al., 2018), which is a PCR-based, short-insert library preparation method designed to maximize historical DNA potential by accounting for low extraction yields. Purification steps were used both at the extraction stage and the library building stage and included both SPRI magnetic beads (Beckman Coulter) and membrane filter MinElute PCR Purification spin columns (Qiagen). Subsequently, Illumina dual indexes (8 bp) were used to facilitate multiplexing of samples. Sequencing was outsourced to NovoGene EU, using the NovaSeq 6,000 with 150 bp paired-end (PE) sequencing.

For the analysis of phylogenetic relationships and monophyly in postglacial northern European *Euphrasia*, sequencing data for 41 individuals were sourced from three previous studies. First, we integrated data for 18 *Euphrasia* samples previously used in a population genomic study of British *Euphrasia*, with a focus on tetraploid species on Fair Isle, Scotland (Becher et al., 2020). This study generated a reference genome of the tetraploid species *E. arctica* (described below) and high-coverage short-read data for 17 additional *Euphrasia* samples. All 18 samples had nuclear SNPs called relative to the reference genome, and plastid genomes and nrDNA arrays assembled *de novo*. Second, low-coverage short-read sequencing data of 12 samples: 10 other British *Euphrasia*, one Austrian *Euphrasia*, and an outgroup *Bartsia alpina*, were available from a study characterizing the landscape of genomic repeats (Becher et al., 2021), with the raw data reanalyzed here and used for *de novo* assembly of plastid genomes and nrDNA, and mapping to the reference genome. Finally, we included short-read data for 11 previously unpublished samples (Twyford, Unpublished Data) where data was available on the Sequence Read Archive (SRA; SRR17976421 - SRR17976431). These represent 10 diverse *Euphrasia* taxa and an outgroup *Neobartsia chilensis*. These low-coverage genome skims were generated from NEB Ultra PCR-based libraries sequenced with 125 bp PE sequencing on the Illumina HiSeq 2500 or 150 bp PE sequencing on the Illumina NovaSeq 6000 at Edinburgh Genomics.

Our final data set included 31 samples collected in Britain, including multiple samples for: *E. anglica* Pugsley (2 samples), *E. arctica* Lange ex Rostr. (7), *E. confusa* Pugsley (2), *E. foulaensis* Towns. ex Wettst. (5), *E. micrantha* Rchb. (7) and *E. vigursii* Davey (2). These also included two putative hybrids (*E. confusa* x *E. foulaensis*, *E. arctica* x *foulaensis*) and two species of putative hybrid origin (*E. rivularis*, *E. vigursii*, Yeo, 1956). We also include a sample of '*Euphrasia fharaidensis*', a UK endemic awaiting formal description (French et al., 2008). Herbarium material from all newly sequenced samples are lodged at E.

Sequence Analysis

Plastid Genome Assembly and Curation

Plastid genomes were assembled for each sample *de novo*, using Novoplasty (Dierckxsens et al., 2017) or GetOrganelle (Jin et al., 2020). Most assemblies were circular, single-contig genomes, however where this was not the case assemblies were subject to additional curation. Specifically, any sample with a large deletion relative to other samples (more than 500 bp) had raw reads mapped back to the *E. arctica* reference plastid genome (Becher et al., 2020) using Geneious v11.1, and with coverage of putative deletions inspected by eye. Several regions for nine samples were then manually added to the plastid genome assemblies. Assembled plastid genomes were manually curated and edited to give a standard order of the large single copy (LSC), inverted repeat (IR), small single copy (SSC), and second copy of the IR, using Geneious. Newly assembled plastid genomes and previous plastids (Becher et al., 2020) were subsequently aligned using MAFFT (Katoh and Standley, 2013).

nrDNA Assembly and Curation

nrDNA arrays were assembled using Novoplasty with the expected assembly size set to 9,000–20,000 bp and using a 1,380 bp seed sequence of the nrDNA cluster, obtained from a run of the RepeatExplorer pipeline (Novák et al., 2013). Variable results were produced by the assembler, with some samples having fully assembled circularized arrays and others having multiple, overlapping contigs. Ambiguous sites were coded with standard nucleotide ambiguity codes, with these sites potentially representing divergent ribotypes maintained within individuals, or uncertainty in the underlying sequencing or assembly. To avoid assembly issues or problems aligning the highly variable external transcribed spacer (ETS), the assemblies were subsequently trimmed to the ~5.8Kb nrDNA coding region (comprising 18S, ITS1, 5.8S, ITS2, 26S, termed the nrDNA array herein). nrDNA arrays were subsequently aligned using MAFFT.

Nuclear Genome Resequencing

Paired sequence reads were mapped to the tetraploid *E. arctica* genome (Becher et al., 2020). This reference genome assembly was produced using high-coverage Illumina data in conjunction with low-coverage Pacific Bioscience data and spans 823 Mb of the ~1.15Gb genome. Characterization of the *E. arctica* genome has shown it to be an old allotetraploid with divergent subgenomes, one of which is closely related to extant British diploid taxa (Becher et al., 2020).

Reads were mapped using the PALEOMIX (Schubert et al., 2014) pipeline, apart from the 18 samples from the study of Becher et al. (2020) which were already aligned and available as a BAM file. The PALEOMIX pipeline is especially designed for the mapping and initial processing of degraded DNA, making it particularly suitable for the herbarium samples included in this study. Raw reads were initially trimmed for ambiguous and low-quality bases at the ends of reads (N, or base quality less than 2). Subsequently, adapter sequences were identified and excised from the 3′ ends of the short reads using AdapterRemoval (v2.2.2) (Schubert et al., 2014). As part of the adapter trimming, read pairs that overlapped by more than 10 bp were merged, and any reads shorter than 25 bp were discarded. The adapter trimmed reads were mapped to the reference genome using the bwa aln algorithm (v0.7.15; Li and Durbin, 2010), with seeds disabled to allow better matches for degraded DNA. Finally, reads aligning to the reference genome with mapping quality less than 30 were discarded from downstream analyses.

Previous analyses of genome resequencing data for 14 British tetraploid and 4 British diploid samples identified a set of 3,454 conserved scaffolds longer than 1 kb, that have coverage consistent with diploid-level mapping depth across all individuals (Becher et al., 2020). In total, these scaffolds represent 46 Mb of the genome. These were proposed to represent disomically inherited nuclear regions homologous across ploidy levels, and belonging to a shared subgenome. Here, we selected these scaffolds for downstream phylogenomic analysis between taxa of varying or unknown ploidy. These scaffolds can be directly compared across ploidy levels and used in conventional phylogenetic packages suitable for diploid taxa, though with the caveat that we may undersample duplicated regions. Sequences for the conserved scaffolds were extracted from the mapping data for each sample based on the scaffold coordinates in the reference genome. FASTA files of consensus sequences were produced with Angsd (Korneliussen et al., 2014) using a quality threshold set at bp-site 3X coverage. Data for each scaffold was filtered using custom Python scripts to remove any samples without representative consensus sequences. Any scaffold with less than three samples represented by consensus sequences were also removed. In addition to focused phylogenomic analyses of the conserved scaffolds, we also investigated analyses of the complete nuclear genome directly from the raw sequence reads (described below).

Phylogenetic Analyses

Phylogenetic analyses were performed independently for plastid genomes, nrDNA arrays, conserved nuclear scaffolds and the complete set of sequence reads. For each data set, phylogenies were annotated with available ploidy and geographic information. Ploidy information came from Metherell and Rumsey (2018) and Becher et al. (2021) for British samples and Gussarova et al. (2008) for other taxa. Ploidy information was available for 35 individuals, with missing information for many newly sequenced non-British samples. Geographic areas were annotated on trees following the areas defined by Gussarova et al. (2008).

Plastid Genome

Phylogenetic analyses of the complete plastid genome sequences were performed using IQTree 2 (Minh et al., 2020) with the best evolutionary model inferred using model fitting and model assessment based on Bayesian Information Criterion. Maximum likelihood trees were obtained using the selected best evolutionary model, and branch support was inferred *via* 1,000 rapid bootstrap replicates. Trees were visualized with FigTree.

nrDNA

Partitioned phylogenetic analysis of the nrDNA arrays was performed in IQTree 2 to account for substitution rate variation (i.e., ITS has a higher substitution rate than other nrDNA regions). Maximum likelihood trees were obtained in IQTree 2, and branch support was inferred *via* 1,000 rapid bootstrap replicates.

Nuclear Genome Resequencing

Phylogenetic analyses were performed on each of the conserved nuclear scaffolds separately, with these then being used to build a putative species tree. IQTree 2 was first used to build Maximum Likelihood trees for each scaffold. Newick Utilities (Junier and Zdobnov, 2010) was used on each IQTree scaffold tree to collapse unsupported branches (bootstrap support, 10) before using Astral III (Zhang et al., 2018) to build a species tree

from the suite of input scaffold trees, with each tree given equal weighting. DiscoVista (Sayyari et al., 2018) was used to compute and visualize the discordance between the species tree and each scaffold tree.

To investigate evolutionary relationships across the genome (not just in the subset of conserved nuclear scaffolds), and to further explore the utility of sequence data from our herbarium samples, we tested MASH (Ondov et al., 2016), an implementation of the MinHash approach to rapidly compute distances between strings. We analyzed the raw sequence reads for the 17 newly sequenced *Euphrasia* herbarium samples using default parameters, with the unrooted neighbor-joining tree visualized using FigTree.

Comparative Phylogenetics

Tanglegrams implemented in Dendroscope version 3.7.5 (Huson and Scornavacca, 2012) were used to detect discordance between the topologies of phylogenetic trees created using different genome partitions. Tanglegrams allow the comparison of rooted trees by rotating nodes to minimize perceived incongruence related to tree visualization. Our comparisons were between: (1) the nrDNA array and the plastid genome, (2) the plastid genome and the species tree from the conserved nuclear genome scaffolds, and (3) the nrDNA array and the species tree from the conserved nuclear genome scaffolds. The tanglegrams were generated with Dendroscope (v 3.7.5) and visualized in R (v 4.1.2) using packages dendextend (v1.15.2), phylogram (v2.1.0) and ape (v5.6-1).

Genomic Analyses of Polyploidy

Previous genomic analyses of British *Euphrasia* inferred individual ploidy, and whether subgenomes are likely to be shared between individuals, based on sequencing coverage per genome scaffold relative to the *E. arctica* reference genome (Becher et al., 2020). In this previous analysis, while all samples had similar coverage across the conserved scaffold set (which are likely to be within a conserved subgenome present across diploid and tetraploid British *Euphrasia*), diploids had no (or very low) coverage in a large number of scaffolds restricted to the tetraploids. Here, we perform a similar mapping depth analysis across our global *Euphrasia* samples to understand whether genome structure is conserved across diverse species. The per-scaffold coverages were computed for each sample, retaining only scaffolds longer than 1 kb. In order to compare the samples within the study and to previously published data (Becher et al., 2020), the per-scaffold coverages were normalized by the average coverage across these scaffolds. Further, the analysis was restricted to ~10,000 scaffolds previously identified as conserved across the genus. Hierarchical clustering was performed on the resulting matrix of normalized per-scaffold coverages at these ~10,000 scaffolds. Coverages were then visualized per sample and per scaffold in a heatmap following Becher et al. (2020). To aid visualization samples were ordered based on relatedness inferred from hierarchical clustering of pairwise Manhattan distances between the samples' mapping depth profiles (Becher et al., 2020).

RESULTS

We were able to recover nuclear genomic data from all 17 herbarium samples, with read counts averaging 276,825,100 per sample (range 241,730,838-312,518,834). These samples were combined with the 42 previously sequenced samples to investigate phylogenomic relationships of plastid genomes, nrDNA arrays, and the nuclear genome.

Plastid Genome Diversity and Phylogenetic Relationships

We successfully assembled the plastid genome for 38 *Euphrasia* individuals and 2 outgroups and compared these to 18 previously assembled plastid genomes. Overall the assembly length was consistent across samples (range: 140,581–145,113 bp for *Euphrasia* species, up to 153,370 bp in *Bartsia alpina*), with a mean size of 144,792 bp across *Euphrasia* species. The average plastid genome GC content was 38.3%. Pairwise identity between *Euphrasia* samples was high at 99.1% and with 87.1% of sites identical across species, with these rising to 99.8 and 98.8%, respectively, across the 31 British samples. The final alignment of 56 *Euphrasia* and 2 outgroup plastid genomes was 157,898 bp in length.

Phylogenetic analyses of plastid genomes revealed a deep split between clades broadly corresponding to Northern Hemisphere taxa, and Southern Hemisphere taxa plus samples from Japan (**Figure 1**). The only exception to this biogeographic split were two samples of the European species *E. cuspidata*, which were placed on a long branch separate from all other samples, consistent with it being a morphological distinct diploid taxon belonging to a separate taxonomic section. Within clades, there was significant phylogenetic complexity, with some patterns of relatedness representing geography, ploidy, or species identities, though many relationships are hard to explain. Within the poorly sampled clade of largely Southern Hemisphere species (represented by eight samples) species do not cluster by geography, with species sampled from the same country (such as New Zealand) separated on the tree.

Within the Northern Hemisphere clade, there is a lack of discernable overall phylogenetic structure and the tree is characterized by extremely short terminal branches. However, some individual or species-level patterns of plastid haplotype sharing and relatedness emerge. These include clusters corresponding to: three samples of *E. micrantha* from Fair Isle; related tetraploid *E. arctica* and *E. foulaensis* from Fair Isle; four diploid samples from *E. anglica* and *E. vigursii* from England; and a cluster of distinct Eurasian species *E. tatarica*, *E. pectinata*, and *E. regelii*. In contrast to these clusters, many more patterns of relatedness appear more complex, for example four other samples of *E. micrantha* (excluding the three samples that cluster) are largely spread throughout the wider Palearctic and Alps group. This pattern of individuals being scattered throughout the Northern Hemisphere clade is also seen with *E. arctica* and *E. foulaensis*.

Overall, the plastid phylogeny highlights the phylogenetic complexity present in the genus, with only weak clustering by

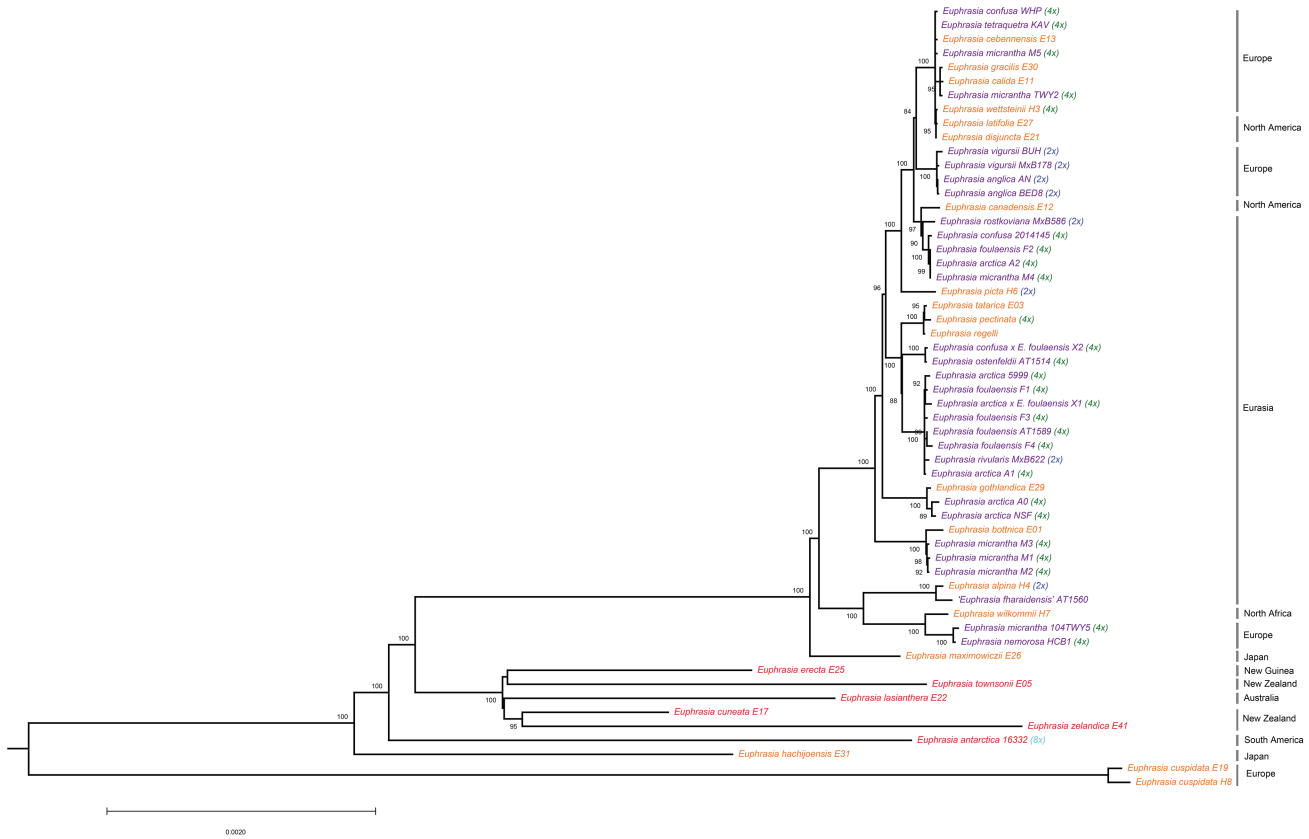


FIGURE 1 | Phylogenetic relationships inferred from plastid genomes for global *Euphrasia* species. Phylogeny constructed based on an alignment of 157,898 sites, using a maximum likelihood analysis implemented in IQTree 2. Branches labeled with bootstrap support values over 85%. Phylogeny rooted on *Bartsia*, which is removed here for better visualization. Tip labels are colored by geography, with red being Southern Hemisphere, purple being UK and orange being European and Asian samples. Tips are also labeled by known ploidy, including diploids (2x, blue), tetraploids (4x, green) and octoploids (8x, light blue).

geography, and with multiple samples within species not showing clear taxonomic coherence.

nrDNA Phylogenetic Relationships

The nrDNA array was successfully assembled for all 38 samples except *E. gothlandica* and *E. micrantha* sample M4, which consistently failed. The new nrDNA assemblies were aligned with the 18 existing assemblies (Becher et al., 2020) to produce a conserved nrDNA alignment of 5,845 bp in length. Per sample conserved nrDNA array lengths varied between 5,571–5,828 bp across the alignment of 55 *Euphrasia* individuals and 2 outgroup samples. Pairwise sequence identity between *Euphrasia* individuals was 99.2, and 91.8% of sites were identical across the alignment. Total GC content was 54.3%. Most sites were unambiguously identified, with only 423 sites in the alignment (0.13% sites) coded as ambiguous. Across samples 21 individuals had no ambiguous sites (36.8%). These samples were either Southern Hemisphere taxa, all of which had no ambiguities (apart from octoploid *E. antarctica*), known Northern Hemisphere diploids (e.g., *E. alpina* and *E. vigursii*), or Northern Hemisphere species of unknown ploidy clustering with the diploids (*E. bottnica*, *E. canadensis*, see below). Exceptions to this finding are diploid *E. rostkoviana* and the putative diploid hybrid species *E. rivularis*, which had some ambiguous sites. All known tetraploid samples from the Northern Hemisphere had at least two ambiguous sites. Most ambiguous sites were not random in their position across the alignment and instead were common at sites segregating for two alleles, indicative of the retention of multiple nrDNA copies rather than assembly errors.

Phylogenetic analyses of the nrDNA array confirmed the presence of a clear and moderately well-supported biogeographic break (BS = 86%) largely corresponding to Northern vs. Southern Hemisphere taxa (**Figure 2**). The southern clade is characterized by long branches typical of older and more divergent lineages, but also reflect artifacts related to poorer sampling. In the Northern Hemisphere clustering is largely by ploidy, with clear separation of most diploid taxa on a well-supported (BS = 100) long branch from most known tetraploids. Within tetraploids, there is a clade comprised of Siberian *E. tatarica* and Chinese *E. regelii* and *E. pectinata*, and a spanish sample of *E. willkommii*. The better-sampled clade of predominantly North Western European tetraploids lacks strong geographic or taxonomic structure, with species and geographic locations being largely intermixed. For example, nrDNA sequences of *E. micrantha* are scattered across the clade and with one sequence on a long branch. However, there is evidence of some geographic

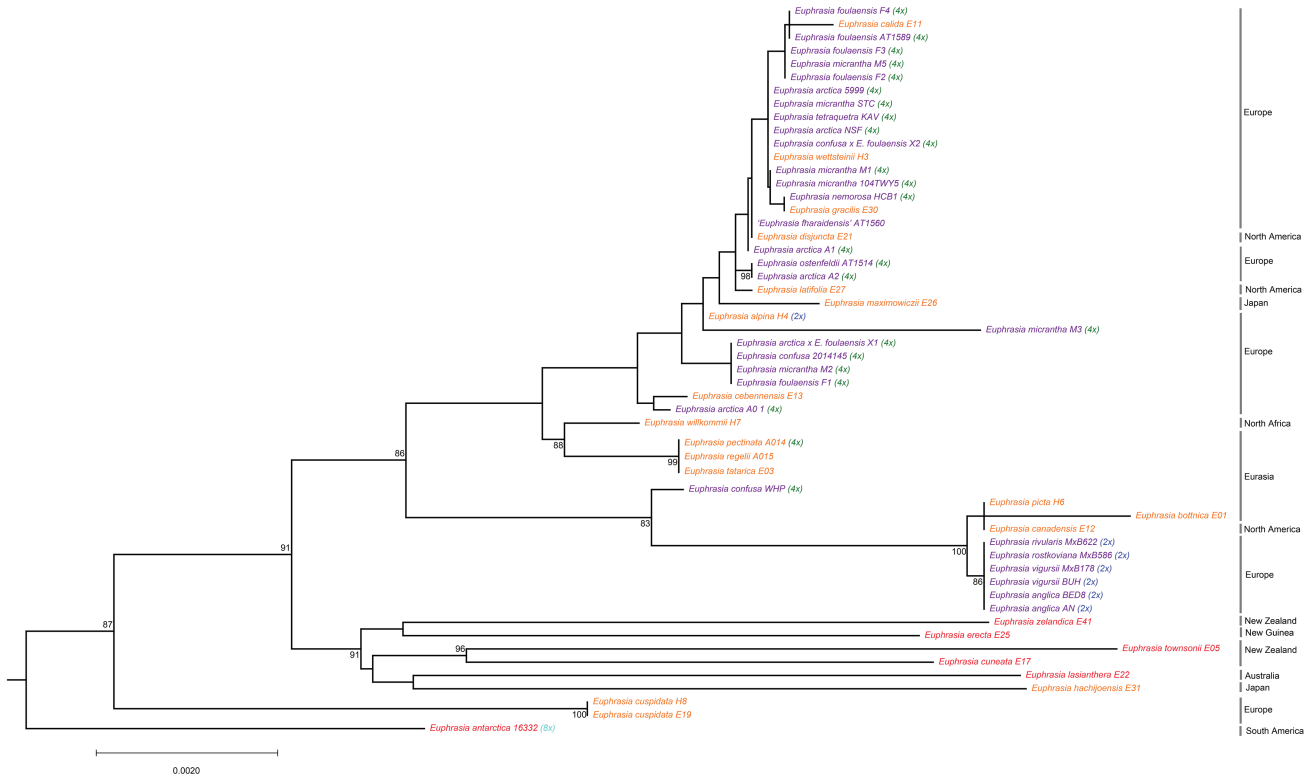


FIGURE 2 | Phylogenetic relationships inferred from nrDNA sequences for global *Euphrasia* species. Phylogeny constructed based on an alignment of 5,845 sites, using a maximum likelihood analysis implemented in IQTree 2. Branches labeled with bootstrap support values over 85%. Phylogeny rooted on *Bartsia*, which is removed here for better visualization. Tip labels are colored and labeled by ploidy, as **Figure 1**.

structure and clustering by taxonomy, for example with four samples representing three species and one hybrid found on Fair Isle possessing identical nrDNA sequences.

Nuclear Genome Resequencing Phylogenetic Relationships

Our species tree of conserved nuclear scaffolds (**Figure 3**), and associated quantification of discordance with DiscoVista (**Figure 4**), revealed that 66% (36/55) of internal nodes had one (and only one) topology across more than 33% of the gene trees. For the remaining (34%) internal branches, there is an equal weighting for a second or third possible topology. Across the species tree, there is notable variation in patterns of discordance. There is generally a lower likelihood of the alternative topology in branches within the clade of Southern Hemisphere taxa and species from Japan (nodes 33–38), which are newly sequenced here, as well as in British diploids (nodes 20–24), or where there is geographically cohesive species sampling of tetraploids (such as three sample of the selfing species *E. foulaensis* sampled on Fair Isle, node 12). In contrast, there is a near equivalent representation of two, or all three, possible relationships in many other branches across the phylogeny, particularly those involving British tetraploids. While the general trend is of greater discordance in branches connecting recent species relationships in European tetraploids, there are also some early diverging nodes with discordance where an alternative topology is frequent (e.g., nodes 40 and 47), showing complexity across the *Euphrasia* phylogeny.

This species tree supports the plastid and nrDNA phylogenies in recovering an early diverging clade predominantly composed of Southern Hemisphere taxa, with the South American taxa placed sister to the Australia-South Asian taxa. The better-sampled Northern Hemisphere clade has the Japanese taxon *E. maximowiczii* included within an otherwise exclusively North American-European clade. Within the Northern Hemisphere clade, there are clearly resolved early diverging relationships between mainland European, Japanese, and North American taxa. British diploid species are resolved as monophyletic, while British tetraploids fall in three large clades, each with a small number of European or North American taxa. Of particular note is one clade where mainland British samples are placed sister to a monophyletic sub-clade of species found on Fair Isle. *Euphrasia micrantha* is found in all three clades including British taxa, including one clade of three *E. micrantha* samples with European *E. gracilis*.

The neighbor-joining tree based on distances estimated by MASH sketches broadly mirrors the topology of the species tree built using trees inferred from the 3,454 conserved sequence scaffolds (**Supplementary Figure S1**), albeit with some short

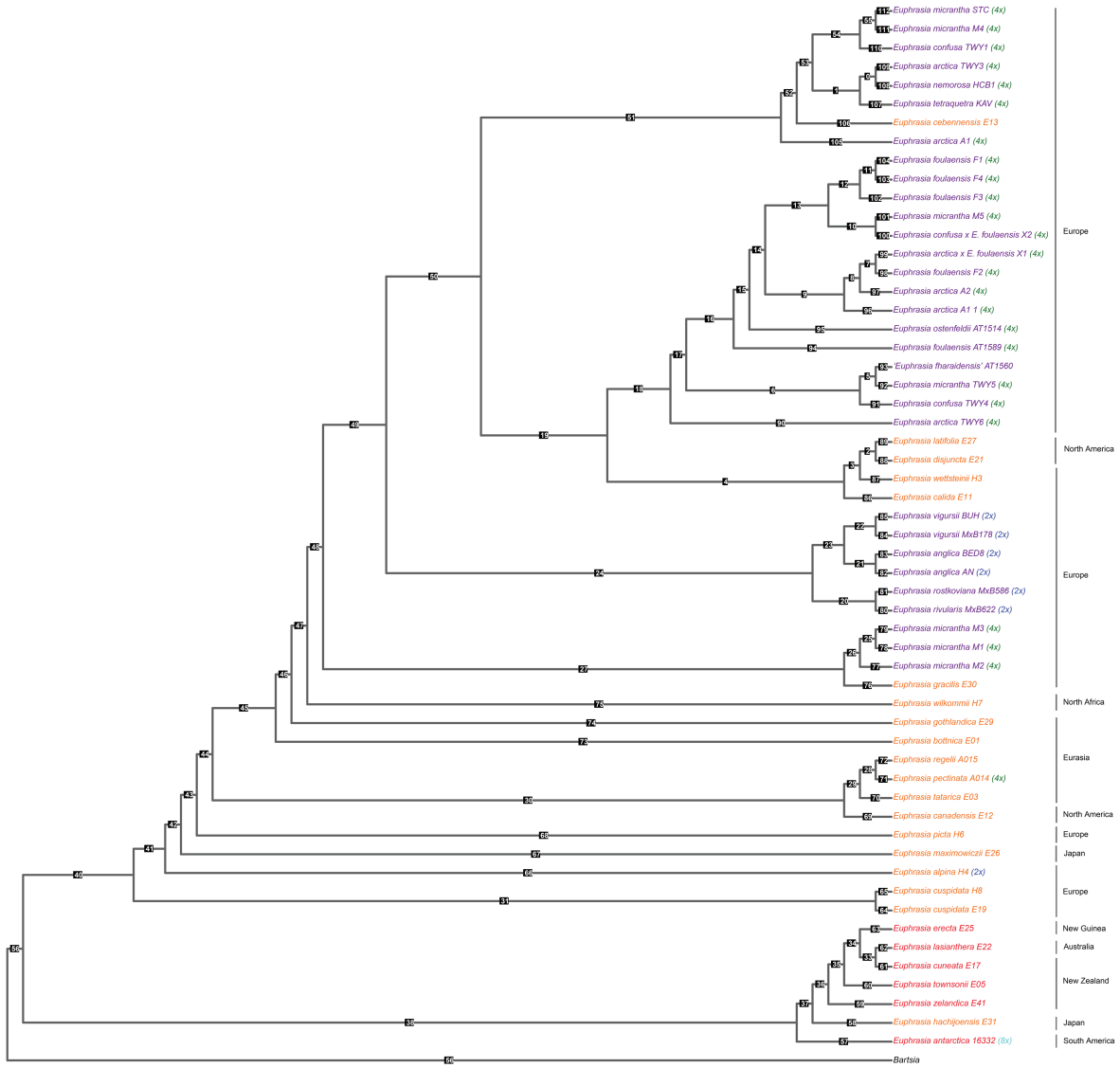


FIGURE 3 | Species tree built from conserved nuclear scaffolds present across *Euphrasia* species. Species tree performed using Astral, based on gene trees for each of 3,454 scaffolds (totaling 46 Mb) generated with IQTree 2. Tips colored by geography and labeled with ploidy as **Figure 1**. Numbered branch labels correspond to separate bar plots quantifying discordance using Discovista, presented in **Figure 4**.

branches. The only sample showing discordant placement relative to the species tree is *Euphrasia gothlandica* (E29), which suffers from low coverage.

Comparative Phylogenetics

Our tanglegram analyses revealed extensive phylogenetic discordance between genomic regions. Given the large amount of informative sites and high support, we focus on comparisons to the nuclear species tree analysis. Incongruence is seen in comparisons with the plastid genome phylogeny (**Figure 5**), as the plastid analysis does not recover any major clades apart from the early diverging predominantly Southern Hemisphere group. Most notably, the plastid phylogeny does not recover a group of diploid taxa or resolve any overall geographic structure within the Northern Hemisphere group (**Figure 5**). When the nuclear species tree is compared to the nrDNA tree (**Figure 6**), there are some similarities including groups corresponding to ploidy, though its placement within the phylogeny differs, appearing on a long branch sister to all Northern Hemisphere tetraploids in the nrDNA tree, and being placed in a more derived clade in the nuclear species tree. Interestingly however, both the nuclear species tree and the plastid tree have some consistent individual-level relationships, such as three samples from Fair Isle, whereas this group is not recovered in the nrDNA analysis. Otherwise the plastid and nrDNA phylogenies are largely incongruent except some early diverging Southern Hemisphere lineages (**Supplementary Figure S2**). One interesting

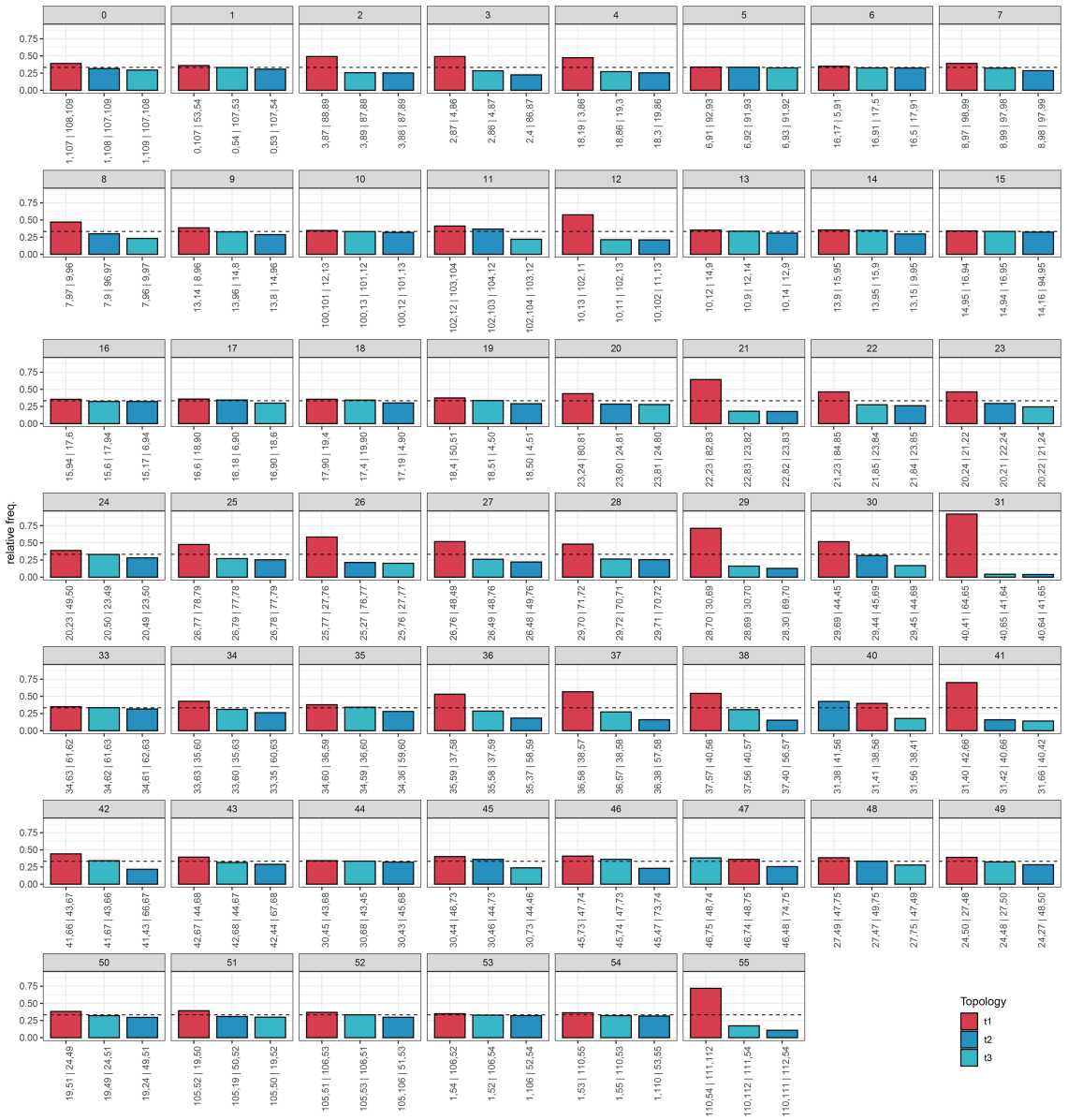


FIGURE 4 | Relative frequencies of alternative tree topologies for *Euphrasia*, computed with DiscoVista. Each bar graph represents potential for the three alternative branching relationships at each focal node labeled in the species tree presented in **Figure 3**. Main topologies are in red, alternative topologies in blue, and the dotted line indicates a 1/3 threshold (equal representation of three topologies). The x-axis is labeled with neighboring branch labels (see Sayyari et al., 2018).

case of incongruence is diploid *E. rivularis*, a species of putative cross-ploidy hybrid origin. This is the only diploid species clustering with a group of tetraploids in the plastid tree, but yet it clusters with other diploids in the nrDNA tree, consistent with its proposed origins.

Genomic Analyses of Polyploidy

We assessed sample ploidy and subgenome relationships based on short-read sequence coverage relative to scaffolds present in the genome of tetraploid *E. arctica*. While the mean mapping depth of ~14X is generally sufficient to infer presence/absence and estimate copy number, there was notable variation, and 5 samples had below 5X mapping which partly obscures patterns (**Supplementary Table S2**).

Euphrasia bottnica (E1, Finland) demonstrated a coverage pattern similar to previously analysed British diploids, with a large set of scaffolds having no (or nearly no) reads mapping, with these absent scaffolds corresponding to the divergent tetraploid subgenome (**Figure 7**). Similarly, *E. calida* (E11, Iceland), *E. cebennensis* (E13, France), *E. disjuncta* (E21, Canada), *E. latifolia* (E27, Canada), and *E. gracilis* (E30, Sweden) show broadly similar coverage patterns to British tetraploids, albeit with higher variance. The other samples had distinctly different mapping depth patterns. However, four samples from the Southern

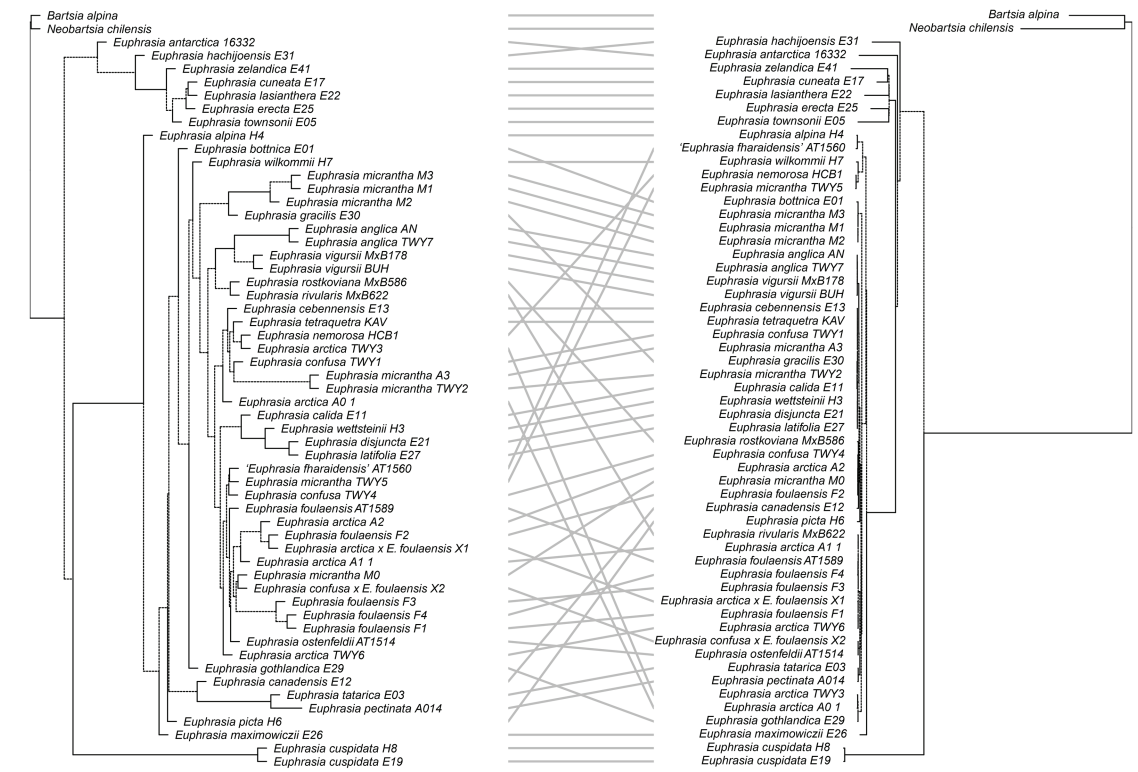


FIGURE 5 | Tanglegram comparing **(A)** the species tree from the conserved nuclear scaffolds and **(B)** the maximum likelihood phylogeny for the plastid genome.

Hemisphere: *E. townsonii* (E5, New Zealand), *E. cuneata* (E17, New Zealand), *E. lasianthera* (E22, Australia), and *E. erecta* (E25, New Guinea), show starkly different coverage patterns for a subset of the conserved scaffolds including both *Euphrasia*-wide conserved scaffolds (3454) and tetraploid-only conserved scaffolds (~7,000). For this subset of scaffolds, these samples show double the average genome coverage, suggesting that these samples are either octoploid or have partial genome duplication post-polyploidization. The inference of polyploid history for samples E3, E12, and E29 is more ambiguous, though these represent tetraploids divergent from the the reference.

DISCUSSION

Taxonomically complex groups are often neglected in genomic studies in preference of more tractable groups with simpler speciation histories. However, the increasing accessibility of genomic sequencing now make it possible to perform comparative genomic analyses in even the most complex plant groups. Here, we perform a phylogenomic analyses of the renowned taxonomically group *Euphrasia*, with a focus on inferring species cohesion and colonization history of British species, and providing first insights into wider genomic variation present across the genus. Combined, these two approaches allow us to consider the role of polyploidy, geography and species barriers in shaping genome-wide variation. Overall we find extensive phylogenomic discordance at both shallow and deep temporal scales, particularly in comparisons involving the plastid genome. Within the postglacial radiation of *Euphrasia* in northern Europe, we detect discrete waves of colonization to Britain from distinct source populations, with complex patterns of individual relatedness that are generally more closely connected to geographic location than species identity. Across our wider *Euphrasia* analyses, we also see strong geographic structure, as well as clustering by ploidy, indicative of reproductive isolation between diploid and tetraploid taxa. Moreover, comparative analyses of sequencing coverage suggest genomic diversity in *Euphrasia* is a consequence of independent evolutionary radiations of tetraploid species. Here, we consider the implications of these results for understanding speciation processes in this enigmatic group, and more widely for understanding genomic variation across diverse *Euphrasia* species.

Extensive Phylogenomic Discordance Across the *Euphrasia* Phylogeny

Phylogenetic discordance has been observed in numerous plant studies and is increasingly considered the norm (Rose et al., 2021). Here, we confirm that signals of phylogenetic discordance observed with Sanger sequencing of few loci at the regional (Wang et al., 2018), and the global scale (Gussarova et al., 2008),

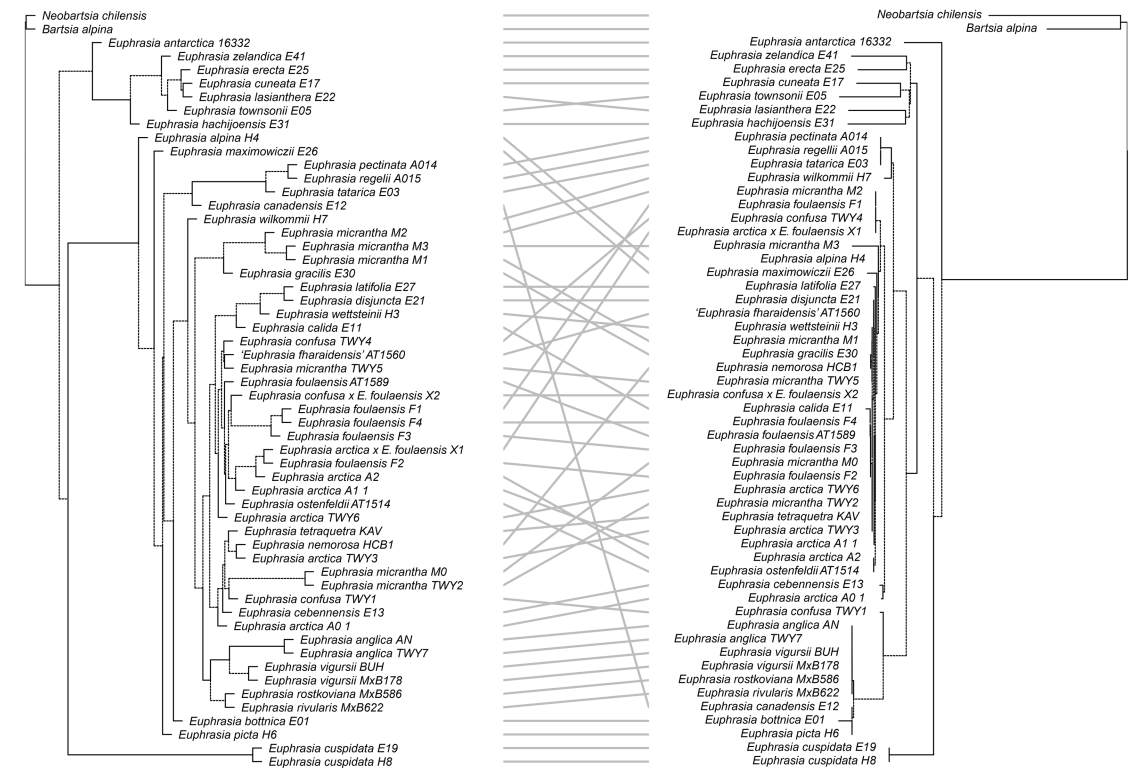


FIGURE 6 | Tanglegram comparing **(A)** the species tree from the conserved nuclear scaffolds and **(B)** the maximum likelihood phylogeny for the nrDNA.

are indeed detected with genomic data. The lack of clear geographic, taxonomic or ploidy-related structuring in the plastid genome phylogeny, coupled with incongruence to the well-supported nuclear genomic phylogeny, shows the plastid does not track other loci and variation is shaped by other evolutionary forces. In particular, frequent hybridization as well as self-fertilization may lead to the geographically restricted fixation of plastid haplotypes giving a local geographic signal conflicting with species boundaries.

While there is less discordance between nrDNA and the nuclear genome there is still conflict, which may be due to the stochasticity underlying a single gene tree, or a specific outcome of nrDNA being maintained in multiple copies and subsequently experiencing concerted evolution (Xu et al., 2017). Perhaps most importantly, nucleotide ambiguity in the tetraploids suggests the maintenance of multiple nrDNA variants following polyploidy, a finding not observed with direct Sanger sequencing in *Euphrasia* (Wang et al., 2018). While intra-individual variation is problematic for reconstructing phylogenies, the broad concordance of major clades between nrDNA and the nuclear species tree suggests this issue does not obscure signal present in the data. Future work will look to clarify evolutionary relationships using phased nrDNA sequences as has been done with polyploid taxa in the Asteraceae (Fehrer et al., 2021), in the hope that this will reveal the currently unknown second progenitor of British tetraploid *Euphrasia* species (Becher et al., 2020; discussed below).

In addition to challenges with the retention of multiple nrDNA copies, there is also ample evidence consistent with incomplete lineage sorting, with individuals with multiple samples combining variation greatly predating speciation, including variants across the tetraploid clade which was previously estimated to be 7.3 million years old (Gussarova et al., 2008). Regardless of the evolutionary processes shaping variation, these results highlight that an "extended barcode" (Coissac et al., 2016) based on plastid and nrDNA from genome skimming will fail to identify species in *Euphrasia*, as well as a range of other complex groups like willows (Wagner et al., 2021). Here, researchers should instead look to sample the nuclear genome, using methods such as whole-genome resequencing or sequence capture with target enrichment probes (Johnson et al., 2019).

Speciation and Colonization History

Our analyses used extensive sampling of British species, coupled with representatives from the geographic range and phylogenetic diversity of *Euphrasia*. This allows us to investigate the colonization and speciation history of recent postglacial species divergence. Previous work has shown that diploid and tetraploid *Euphrasia* diverged long before recent pleistocene glaciation

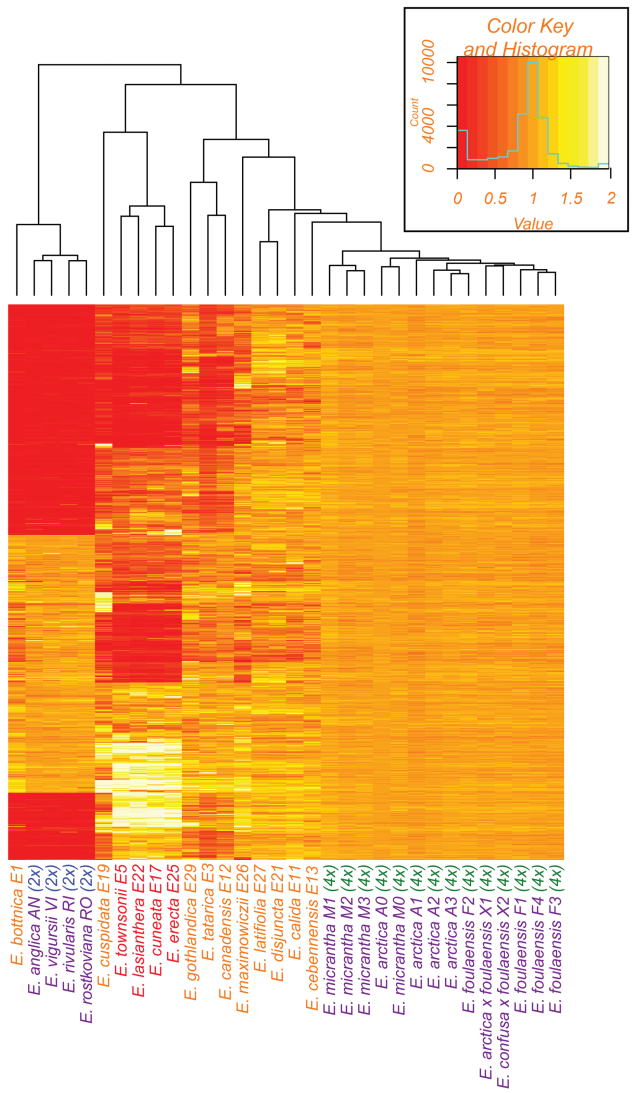


FIGURE 7 | Heatmap of relative mapping depth of *Euphrasia* whole-genome resequencing data relative to the *E. arctica* reference genome. Blocks of color represent mapping coverage (see key, top right) for each scaffold along the y-axis. Species names are colored by geographic area and ploidy as in **Figure 1**. The tree is based on hierarchical clustering of pairwise Manhattan distances between the samples' mapping depth profiles.

events and colonized Britain independently (Wang et al., 2018). Our higher resolution nuclear genomic data further show British tetraploids are present in at least three clades mixed with mainland European samples, and therefore likely represent at least three waves of colonization. These results mirror many other plant phylogeographic studies, where different genetic lineages present in Britain have colonized recurrently from continental Europe and co-exist (Wood et al., 2018).

In addition to broad-scale phylogeographic patterns, our work also reveals fine-scale insights, such as showing that the geographically widespread small-flowered selfing taxon *Euphrasia micrantha* is polyphyletic in analyses of the nuclear genome, plastid and nrDNA array. This finding is surprising, as it is one of the most morphologically distinct *Euphrasia* species, characterized by purple leaves, stems, and flowers, and has a distinctive ecology being predominantly found in heather moorland (Stone, 2012). Our more limited sampling of a number of other taxa reveals a similar lack of species cohesion. These results are in line with previous genetic and genomic studies showing genetic variation in *Euphrasia* often clusters by geography rather than by species, at least within a ploidy level (French et al., 2008; Becher et al., 2020). While phenotypic plasticity and taxonomic confusion are profound challenges for studies of *Euphrasia* (Brown et al., 2020), these are of limited concern at least for *E. micrantha*, which maintains its morphological distinctiveness under a range of conditions and is unlikely to be confused by *Euphrasia* experts. This leaves a number of non-mutually exclusive explanations underlying the origin and maintenance of the species.

Firstly, our analyses here either looked at largely non-recombinant single genomic regions (plastid or nrDNA), or aggregated regions with partially independent evolutionary histories (conserved nuclear scaffolds). In particular the use of scaffolds conserved across individuals means we have only investigated genomic relatedness in one subgenome of the tetraploid. As such these results may have overlooked or masked more subtle genomic signatures at individual nuclear loci. It may be that *E. micrantha* and other *Euphrasia* species are monophyletic at specific nuclear regions underlying species differences that were not analyzed separately here, but experience homogenizing gene flow, such as *via* hybridization across the rest of the genome (Twyford and Friedman, 2015). This remains a distinct possibility given weak reproductive barriers between *Euphrasia* species. Such an explanation would be consistent with either a single origin of the species followed by hybridization, or multiple origins at different sites perhaps from a shared pool of standing genetic variation (i.e., combinatorial speciation *sensu* Marques et al., 2019). However, the maintenance of such high genetic diversity and divergent haplotypes within species, and within a single sampling location, particularly in a selfing taxon, may also point toward the presence of cryptic species. We are currently pursuing these hypotheses using further genomic sequencing of population samples, where we aim to quantify the extent of hybridization between individuals from different geographic areas, and with contrasting ploidy and mating systems.

Genome Evolution and Polyploidy

The previous study of Becher et al. (2020) identified an allotetraploid origin of British *Euphrasia*, with one subgenome closely related to British diploids. Our analysis of sequencing coverage revealed that *Euphrasia bottnica* sampled from Finland possesses a similar genome to extant British diploids, and this species clusters with diploids in the nrDNA phylogeny, suggesting a shared genomic affinity across this region of postglacial recolonization. Similarly, a number of European tetraploid taxa, such as *E. calida* and *E. cebennensis*, as well as Canadian *E. disjuncta* and *E. latifolia*, have similar sequence coverage patterns to British tetraploids, consistent with a

shared allopolyploid origin for these Northern Hemisphere taxa. Perhaps more notable are cases such as *E. cuneata* and *E. townsonii* (New Zealand), *E. lasianthera* (Australia), *E. erecta* (New Guinea), and *E. cuspidata* (Austria), which are characterized by an extremely different coverage patterns, with some scaffolds in the *E. arctica* reference having no coverage and others having two-fold coverage. Firm conclusions of the ploidy and subgenome constituents of these taxa are hard to make given they are likely to show substantial divergence from the British *E. arctica* reference genome, and are also characterized by low mapping coverage. We have attempted to further characterize sequence reads from these individuals using k-mer based approaches including KAT and Tetmer (Mapleson et al., 2017; Becher et al., 2020), but failed to retrieve a clear signal (Unpublished Results), likely due to low sequencing coverage and potential DNA degradation from these herbarium specimens. Regardless, the finding that these divergent species possess a different genome structure to other *Euphrasia* warrants further study, and suggests recurrent polyploidy in the genus, in line with known tetraploids and hexaploids being scattered across the *Euphrasia* phylogeny.

Prospect of Phylogenomic Analyses in Taxonomically Complex Groups

Taxonomically complex groups frequently pose the joint challenges of taxonomic issues, where species definitions may be uncertain and monographic work is often sorely needed, and systematic/phylogenetic issues, where molecular phylogenies show complex patterns of relatedness. Both issues are relevant to *Euphrasia*. Much has been learnt about species limits since the world monograph of the genus by von Wettstein (1896) and the European revision by Yeo (1978), not least the extent of phenotypic plasticity that taxa exhibit in response to host species and ecological conditions (Karlsson, 1984; Zopfi, 1997; Brown et al., 2020). Our general view is that *Euphrasia* species as currently described, particularly in Britain, may have been too finely divided, and future monographic work may look to "lump" a range of species where trait differences are minimal or prove unreliable. Despite these taxonomic issues, we note that even the most distinct species, such as *E. micrantha*, are not monophyletic (discussed above), showing phylogenetic complexity will persist even following taxonomic realignment of species. This is unsurprising given the nature of these species, showing recent speciation, rampant hybridization, and selfing or mixed-mating systems.

One source of samples that has proved particularly useful in our study has been verified material present in herbarium collections. The search for genomic tools that reliably recover information from herbarium specimens is driven by the incredible amount of historic plant diversity contained within these collections (Särkinen et al., 2012; Buerki and Baker, 2016; Brewer et al., 2019). Accessing herbaria's genomic data will allow researchers to (figuratively) travel through time and space to study extinct taxa and changes in genetic diversity over time. Many studies have now demonstrated the efficacy of genome skimming or target capture to recover genomic data from historical samples (Zeng et al., 2018). Both these approaches rely on a form of "enrichment," with genome skimming analyzing "naturally enriched" regions (i.e., those at high copy number) while target capture enriches regions homologous to target baits. Here, we show that non-enriched, direct whole-genome sequencing can be successfully used for degraded herbarium material. As well as being used to assemble the plastid and nrDNA array, we were able to map sufficient reads to the *E. arctica* reference genome to infer sample ploidy. While useful, however, there were notable issues with these analyses, particularly due to low sample mapping depths. This may either be a consequence of species divergence or contamination, with previous work showing over 70% of sequence reads from herbarium material may be contaminants (Bieker et al., 2020). This issue, combined with DNA error profiles of dried plant tissue, prevented us performing further characterization of these genomes, and suggests future work must oversample herbarium DNA to ensure sufficient data post bioinformatic filtering, or use silica dried plant tissue where available (Brewer et al., 2019). Despite these concerns, we found neighbor-joining trees generated from raw sequence reads from herbarium samples, inferred using MASH, mirrored the topology of our more rigorous scaffold-based nuclear phylogenetic analyses, suggesting sample degradation and contamination do not obscure the main signal of genome-wide relatedness.

Phylogenomic analyses of taxonomically complex groups are often made difficult due to reticulation coupled with polyploidy. Here, we circumvented a number of issues by analyzing haploid plastid genomes, though our hope that nrDNA would have been homogenized within an individual appears not to be the case. We similarly focused our nuclear genome analyses on conserved disomically inherited scaffolds, allowing us to compare across diverse ploidies and to represent evolutionary relationships using a species tree analysis. We chose not to further interrogate genomic relationships within putative subgenomes due to the uncertain homology across these diverse species of differing ploidy and with potentially different parental progenitors. Future work in *Euphrasia*, and other taxonomically complex groups, may look to long-read sequencing and pangenome analyses to better represent structural genomic variation across diverse taxa without reference bias, and to provide robust sorting of homoeologs between subgenomes (Bayer et al., 2021). More integrated polyploidy-aware phylogenomic networks, such as alloPPnet, are also likely to prove fruitful, particularly in the future if this or other methods are developed that are less computationally demanding and allow larger multi-sample data sets as well as more diverse ploidy levels (Rothfels, 2021).

Conclusion

Studies of the extent of discordance in phylogenies have given important insights into a range of topics, including hybridization (Patterson et al., 2012; Martin et al., 2013) and hybrid speciation (Köhler et al., 2021), evolutionary conflict (Hedtke and Hillis, 2010),

horizontal gene transfer (Davis et al., 2005) and rates of phenotypic innovation (Parins-Fukuchi et al., 2021). Our study shows taxonomically complex *Euphrasia* represent a genus where phylogenetic discordance is extensive, at both shallow and deep nodes in the phylogeny. This discordance is likely to be driven by the interaction of different processes, including recurrent rounds of polyploidy, rampant hybridization, and recent postglacial species divergence. Future work will look to estimate the contribution of these processes to phylogenetic conflict in chromosome level genome assemblies.

DATA AVAILABILITY STATEMENT

The newly generated raw sequence reads are available in the SRA, and plastid genomes and nrDNA sequences are in Genbank. The sequence alignments, phylogenetic trees and scripts for Tanglegrams are provided in Dryad (https://doi.org/10.5061/dryad.jh9w0vtd).

AUTHOR CONTRIBUTIONS

PG, AT, CP, and RN designed the research. GG and AT provided samples. PG generated sequencing data. PG, HB, SG, and AT analyzed the data. PG and AT wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the NERC International Opportunities Fund Grant NE/N006739/1 "Evolutionary consequences of facultative plant parasitism" awarded to AT, GG, CP, and RN, and grants NE/R010609/1 and NE/L011336/1 awarded to AT. Plant. ID has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 765000. The Royal Botanic Garden Edinburgh is supported by the Scottish Government's Rural and Environmental Science and Analytical Services Division.

ACKNOWLEDGMENTS

We thank Chris Metherell for help with plant identification, Edgar Wong for assistance with library making for British samples, Max Brown for providing sequences of British diploid *Euphrasia species*, and Julia Naumann for help with plastid genome curation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.869583/full#supplementary-material

REFERENCES

Bakker, F. T., Lei, D., Yu, J., Mohammadin, S., Wei, Z., van de Kerke, S., et al. (2016). Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. *Biol. J. Linn. Soc.* 117, 33–43. doi: 10.1111/bij.12642

Bayer, P. E., Scheben, A., Golicz, A. A., Yuan, Y., Faure, S., Lee, H., et al. (2021). Modelling of gene loss propensity in the pangenomes of three *Brassica* species suggests different mechanisms between polyploids and diploids. *Plant Biotechnol. J.* 19, 2488–2500. doi: 10.1111/pbi.13674

Becher, H., Brown, M. R., Powell, G., Metherell, C., Riddiford, N. J., and Twyford, A. D. (2020). Maintenance of species differences in closely related tetraploid parasitic *Euphrasia* (Orobanchaceae) on an isolated island. *Plant Comms.* 1:100105. doi: 10.1016/j.xplc.2020.100105

Becher, H., Powell, R. F., Brown, M. R., Metherell, C., Pellicer, J., Leitch, I. J., et al. (2021). The nature of intraspecific and interspecific genome size variation in taxonomically complex eyebrights. *Ann. Bot.* 128, 639–651. doi: 10.1093/aob/mcab102

Bieker, V. C., Barreiro, F. S., Rasmussen, J. A., Brunier, M., Wales, N., and Martin, M. D. (2020). Metagenomic analysis of historical herbarium specimens reveals a postmortem microbial community. *Mol. Ecol. Res.* 20, 1206–1219. doi: 10.1111/1755-0998.13174

Brandrud, M. K., Baar, J., Lorenzo, M. T., Athanasiadis, A., Bateman, R. M., Chase, M. W., et al. (2020). Phylogenomic relationships of diploids and the origins of allotetraploids in *Dactylorhiza* (Orchidaceae). *Syst. Biol.* 69, 91–109. doi: 10.1093/sysbio/syz035

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., et al. (2019). Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Front. Plant Sci.* 10:1102. doi: 10.3389/fpls.2019.01102

Brown, M. R., Frachon, N., Wong, E. L. Y., Metherell, C., and Brown, M. R. (2020). Life history evolution and phenotypic plasticity in parasitic eyebrights (*Euphrasia*, Orobanchaceae). *Am. J. Bot.* 107, 456–465. doi: 10.1002/ajb2.1445

Brown, M. R., Moore, P. G. P., and Twyford, A. D. (2021). Performance of generalist hemiparasitic *Euphrasia* across a phylogenetically diverse host spectrum. *New Phytol.* 232, 2165–2174. doi: 10.1111/nph.17752

Buerki, S., and Baker, W. J. (2016). Collections-based research in the genomic era. *Biol. J. Linn. Soc.* 117, 5–10.

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., et al. (2018). Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9, 410–419. doi: 10.1111/2041-210X.12871

Chen, Z. J., Sreedasyam, A., Ando, A., Song, Q., De Santiago, L. M., Hulse-Kemp, A. M., et al. (2020). Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. *Nat. Genet.* 52, 525–533. doi: 10.1038/s41588-020-0614-5

Coissac, E., Hollingsworth, P. M., Lavergne, S., and Taberlet, P. T. (2016). From barcodes to genomes: extending the concept of DNA barcoding. *Mol. Ecol.* 25, 1423–1428. doi: 10.1111/mec.13549

Davis, C. C., Anderson, W. R., and Wurdack, K. J. (2005). Gene transfer from a parasitic flowering plant to a fern. *Proc. Royal. Soc. B.* 272, 2237–2242. doi: 10.1098/rspb.2005.3226

Dierckxsens, N., Mardulyn, P., and Smits, G. S. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45:e18. doi: 10.1093/nar/gkw955

Ennos, R. A., French, G. C., and Hollingsworth, P. M. (2005). Conserving taxonomic complexity. *Trends Ecol. Evol.* 20, 164–168. doi: 10.1016/j.tree.2005.01.012

Fehrer, J., Slavíková, R., Paštová, L., Josefiová, J., Mráz, P., Chrtek, J., et al. (2021). Molecular evolution and organization of ribosomal dna in the hawkweed tribe Hieraciinae (Cichorieae, Asteraceae). *Front. Plant Sci.* 12:647375. doi: 10.3389/fpls.2021.647375

French, G. C., Ennos, R. A., Silverside, A. J., and Hollingsworth, P. M. (2005). The relationship between flower size, inbreeding coefficient and inferred selfing rate in British *Euphrasia* species. *Heredity* 94, 44–51. doi: 10.1038/sj.hdy.6800553

French, G. C., Hollingsworth, P. M., Silverside, A. J., and Ennos, R. A. (2008). Genetics, taxonomy and the conservation of British *Euphrasia*. *Conserv. Genet.* 9, 1547–1562. doi: 10.1007/s10592-007-9494-9

Gussarova, G., Popp, M., Vitek, E., and Brochmann, C. (2008). Molecular phylogeny and biogeography of the bipolar *Euphrasia* (Orobanchaceae): recent radiations in an old genus. *Mol. Phylogenet. Evol.* 48, 444–460. doi: 10.1016/j.ympev.2008.05.002

Hedtke, S. M., and Hillis, D. M. (2010). The potential role of androgenesis in cytoplasmic–nuclear phylogenetic discordance. *Syst. Biol.* 60, 87–96. doi: 10.1093/sysbio/syq070

Huson, D. H., and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* 61, 1061–1067. doi: 10.1093/sysbio/sys062

Jin, J. J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 21, 241. doi: 10.1186/s13059-020-02154-5

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigue, L. R., Cowan, R. S., Devault, A., et al. (2019). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* 68, 594–606. doi: 10.1093/sysbio/syy086

Junier, T., and Zdobnov, E. M. (2010). The newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinform.* 26, 1669–1670. doi: 10.1093/bioinformatics/btq243

Karlsson, T. (1984). Early–flowering taxa of *Euphrasia* (Scrophulariaceae) on Gotland, Sweden. *Nord. J. Bot.* 4, 303–326. doi: 10.1111/j.1756-1051.1984.tb01502.x

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15:356. doi: 10.1186/s12859-014-0356-4

Köhler, M., Oakley, L. J., Font, F., Las Peñas, M. L., and Majure, L. C. (2021). On the continuum of evolution: a putative new hybrid speciation event in *Opuntia* (Cactaceae) between a native and an introduced species in southern South America. *Syst. Biodivers.* 19, 1026–1039. doi: 10.1080/14772000.2021.1967510

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., and Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). *New Phytol.* 221, 515–526. doi: 10.1111/nph.15386

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinform.* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Mapleson, D., Accinelli, G. G., Kettleborough, G., Wright, J., and Clavijo, B. J. (2017). KAT: a k-mer analysis toolkit to quality control ngs datasets and genome assemblies. *Bioinformatics.* 33, 574–576. doi: 10.1093/bioinformatics/btw663

Marques, D. A., Meier, J. I., and Seehausen, O. (2019). A combinatorial view on speciation and adaptive radiation. *Trends Ecol. Evol.* 34, 531–544. doi: 10.1016/j.tree.2019.02.008

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., et al. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* 23, 1817–1828. doi: 10.1101/gr.159426.113

Metherell, C., and Rumsey, F. J. (2018). *Eyebrights (Euphrasia) of the UK and Ireland*. Durham: Botanical Society of Britain and Ireland.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Haeseler, A. V., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., and Macas, J. (2013). RepeatExplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinform.* 29, 792–793. doi: 10.1093/bioinformatics/btt054

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x

Parins-Fukuchi, C., Stull, G. W., and Smith, S. A. (2021). Phylogenomic conflict coincides with rapid morphological innovation. *Proc. Natl. Acad. Sci. U. S. A.* 118:e2023058118. doi: 10.1073/pnas.2023058118

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093. doi: 10.1534/genetics.112.145037

Rieseberg, L. H., and Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trends Plants* 5, 65–84.

Rose, J. P., Toledo, C. A. P., Lemmon, E. M., Lemmon, A. R., and Sytsma, K. J. (2021). Out of sight, Out of mind: widespread nuclear and plastid-nuclear discordance in the flowering plant genus *Polemonium* (polemoniaceae) suggests widespread historical gene flow despite limited nuclear signal. *Syst. Biol.* 70, 162–180. doi: 10.1093/sysbio/syaa049

Rothfels, C. J. (2021). Polyploid phylogenetics. *New Phytol.* 230, 66–72. doi: 10.1111/nph.17105

Sayyari, E., Whitfield, J. B., and Mirarab, S. (2018). DiscoVista: interpretable visualizations of gene tree discordance. *Mol. Phylogenet. Evol.* 122, 110–115. doi: 10.1016/j.ympev.2018.01.019

Särkinen, T., Staats, M., Richardson, J. E., Cowan, R. S., and Bakker, F. T. (2012). How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS ONE* 7:e43808. doi: 10.1371/journal.pone.0043808

Schubert, M., Ermini, L., Sarkissian, C. D., Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by snp detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9, 1056–1082. doi: 10.1038/nprot.2014.063

Squirrell, J., Hollingsworth, P. M., Bateman, R. M., Tebbitt, M. C., and Hollingsworth, M. L. (2002). Taxonomic complexity and breeding system transitions: conservation genetics of the *Epipactis leptochila* complex (Orchidaceae). *Mol. Ecol.* 11, 1957–1964. doi: 10.1046/j.1365-294X.2002.01610.x

Stace, C. A., Preston, C. D., and Pearman, D. A. (2015). *Hybrid Flora of the British Isles*. Bristol: Botanical Society of Britain and Ireland.

Stone, H. (2012). *The Evolution and Conservation of Tetraploid Euphrasia L. in Britain*. Edinburgh: University of Edinburgh.

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., and Liston, A. (2012). Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *Am. J. Bot.* 99, 349–364. doi: 10.3732/ajb.1100335

Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A., and Smith, S. A. (2020). Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. *Am. J. Bot.* 107, 790–805. doi: 10.1002/ajb2.1468

Twyford, A. D., and Friedman, J. (2015). Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion. *Evolution* 69, 1476–1486. doi: 10.1111/evo.12663

von Wettstein, R. (1896). *Monographie der gattung Euphrasia*. Leipzig: Verlag von Wilhelm Engelmann

Wagner, N. D., Volf, M., and Hörandl, E. (2021). Highly diverse shrub willows (*Salix* l.) share highly similar plastomes. *Front. Plant Sci.* 12:662715. doi: 10.3389/fpls.2021.662715

Wang, X., Gussarova, G., Ruhsam, M., de Vere, N., Metherell, C., Hollingsworth, P. M., et al. (2018). DNA barcoding a taxonomically complex hemiparasitic genus reveals deep divergence between ploidy levels but lack of species-level resolution. *AoB Plants* 10:ly026. doi: 10.1093/aobpla/ply026

Wood, D. P., Olofsson, J. K., McKenzie, S. W., and Dunning, L. T. (2018). Contrasting phylogeographic structures between freshwater lycopods and angiosperms in the British Isles. *Bot. Lett.* 165, 476–486. doi: 10.1080/23818107.2018.1505545

Xu, B., Zeng, X.-M., Gao, X.-F., Jin, D.-P., and Zhang, L.-B. (2017). ITS non-concerted evolution and rampant hybridization in the legume genus *Lespedeza* (Fabaceae). *Sci. Rep.* 7:40057. doi: 10.1038/srep40057

Yeo, P. F. (1956). Hybridization between diploid and Tetraploid species of *Euphrasia*. *Watsonia* 3, 253–269.

Yeo, P. (1964). The growth of *Euphrasia* in cultivation. *Watsonia* 6, 1–24.

Yeo, P. F. (1978). A taxonomic revision of *Euphrasia* in Europe. *Bot. J. Linn. Soc.* 77, 223–334. doi: 10.1111/j.1095-8339.1978.tb01401.x

Zeng, C.-X., Hollingsworth, P. M., Yang, J., He, Z.-S., Zhang, Z.-R., Li, D.-Z., et al. (2018). Genome skimming herbarium specimens for DNA barcoding and phylogenomics. *Plant Methods* 14:43. doi: 10.1186/s13007-018-0300-0

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

Zopfi, H. J. (1997). Ecotypic variation of *Euphrasia rostkoviana* Hayne (Scrophulariaceae) in relation to grassland management. *Flora* 192, 279–295. doi: 10.1016/S0367-2530(17)30793-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 September 2022
DOI 10.3389/fpls.2022.967456

OPEN ACCESS

EDITED BY
Stefan Wanke,
Technical University Dresden, Germany

REVIEWED BY
Carolina Granados Mendoza,
National Autonomous University of Mexico, Mexico
Diego F. Morales-Briones,
Ludwig Maximilian University of Munich, Germany

*CORRESPONDENCE
Olle Thureborn
olle.thureborn@su.se

SPECIALTY SECTION
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

RECEIVED 12 June 2022
ACCEPTED 03 August 2022
PUBLISHED 08 September 2022

CITATION
Thureborn O, Razafimandimbison SG, Wikström N and Rydin C (2022) Target capture data resolve recalcitrant relationships in the coffee family (Rubioideae, Rubiaceae). *Front. Plant Sci.* 13:967456. doi: 10.3389/fpls.2022.967456

Target capture data resolve recalcitrant relationships in the coffee family (Rubioideae, Rubiaceae)

Olle Thureborn[1]*, Sylvain G. Razafimandimbison[2], Niklas Wikström[1,3] and Catarina Rydin[1,3]

[1]Department of Ecology, Environment and Plant Sciences, Stockholm University, Stockholm, Sweden, [2]Department of Botany, Swedish Museum of Natural History, Stockholm, Sweden, [3]Bergius Foundation, Royal Swedish Academy of Sciences, Stockholm, Sweden

Subfamily Rubioideae is the largest of the main lineages in the coffee family (Rubiaceae), with over 8,000 species and 29 tribes. Phylogenetic relationships among tribes and other major clades within this group of plants are still only partly resolved despite considerable efforts. While previous studies have mainly utilized data from the organellar genomes and nuclear ribosomal DNA, we here use a large number of low-copy nuclear genes obtained via a target capture approach to infer phylogenetic relationships within Rubioideae. We included 101 Rubioideae species representing all but two (the monogeneric tribes Foonchewieae and Aitchinsonieae) of the currently recognized tribes, and all but one non-monogeneric tribe were represented by more than one genus. Using data from the 353 genes targeted with the universal Angiosperms353 probe set we investigated the impact of data type, analytical approach, and potential paralogs on phylogenetic reconstruction. We inferred a robust phylogenetic hypothesis of Rubioideae with the vast majority (or all) nodes being highly supported across all analyses and datasets and few incongruences between the inferred topologies. The results were similar to those of previous studies but novel relationships were also identified. We found that supercontigs [coding sequence (CDS) + non-coding sequence] clearly outperformed CDS data in levels of support and gene tree congruence. The full datasets (353 genes) outperformed the datasets with potentially paralogous genes removed (186 genes) in levels of support but increased gene tree incongruence slightly. The pattern of gene tree conflict at short internal branches were often consistent with high levels of incomplete lineage sorting (ILS) due to rapid speciation in the group. While concatenation- and coalescence-based trees mainly agreed, the observed phylogenetic discordance between the two approaches may be best explained by their differences in accounting for ILS. The use of target capture data

greatly improved our confidence and understanding of the Rubioideae phylogeny, highlighted by the increased support for previously uncertain relationships and the increased possibility to explore sources of underlying phylogenetic discordance.

KEYWORDS

Angiosperms353, incomplete lineage sorting, non-coding DNA, nuclear phylogeny, phylogenomics, Rubiaceae, Rubioideae, target capture

Introduction

The subfamily Rubioideae, the largest of the major lineages of the species-rich and morphologically diverse coffee family (the Rubiaceae), includes over 8,000 species (Wikström et al., 2020). The members of the subfamily are characterized as herbs or shrubs (rarely trees) with tissues containing raphides (calcium oxalate crystals), valvate corolla aestivation, indumentum of septate hairs and heterostylous flowers (e.g., Robbrecht, 1988; Bremer and Manen, 2000; Robbrecht and Manen, 2006; Bremer and Eriksson, 2009). As for the remaining family, most species are found in tropical and subtropical regions around the world, however, several species of the tribes Anthospermeae, Putorieae, Rubieae, and Theligoneae are distributed in temperate regions. The wind-pollinated flowers in the tribes Anthospermeae and Theligoneae are also an unusual trait, relative to other Rubiaceae, found in this subfamily. The four aforementioned temperate tribes belong to one of the major clades within the subfamily, the cosmopolitan and mainly herbaceous Spermacoceae alliance, which contain over 3,000 species. Together the tribes Spermacoceae and Rubieae make up the bulk of species with more than 1,300 and 900 species, respectively (Wikström et al., 2020). The other major informal group of Rubioideae, the pan-tropical and mainly woody Psychotrieae alliance, also contains over 3,000 species, of which most belong to the tribes Palicoureeae and Psychotrieae, much due to the large genera *Psychotria* and *Palicourea*, with about 1,600 and 800 species, respectively (Razafimandimbison et al., 2008, 2014; Davis et al., 2009).

In total, Wikström et al. (2020) recognized 27 tribes in the subfamily Rubioideae in their summary, based on previous molecular phylogenetic studies. Recently two additional monospecific tribes have been described; the tribe Seychelleeae, which is sister to the tribe Colletoecemateae (Razafimandimbison et al., 2020), and the tribe Aitchinsonieae, which is placed in the Putorieae-Rubieae-Theligoneae clade (also referred to as the Rubieae complex, Bordbar et al., 2021). The Rubioideae thus include the two major groups the Psychotrieae and the Spermacoceae alliances, and seven additional tribes: Colletoecemateae, Seychelleeae, Urophylleae, Ophiorrhizeae, Lasiantheae, Perameae, and Coussareeae. The members of the Psychotrieae alliance are classified in nine tribes: Craterispermeae, Gaertnereae, Mitchelleae, Morindeae, Palicoureeae, Prismatomerideae, Psychotrieae, Schizocoleeae, and Schradereae. In the Spermacoceae alliance, 13 tribes are recognized: Aitchinsonieae, Argostemmateae, Anthospermeae, Cyanoneuroneae, Danaideae, Dunnieae, Foonchewieae, Knoxieae, Paederieae, Putorieae, Rubieae, Spermacoceae, and Theligoneae.

Until recent years, phylogenetic studies in the Rubioideae have mainly relied on information from selected plastid markers (e.g., *atpB-rbcL, rbcL, rps16, trnT-trnL-trnF, ndhF*) (Andersson and Rova, 1999; Bremer and Manen, 2000; Piesschaert et al., 2000; Robbrecht and Manen, 2006; Rydin et al., 2008; Bremer and Eriksson, 2009; Wikström et al., 2015; Janssens et al., 2016) or plastid markers combined with a few nuclear ribosomal regions (e.g., nrITS and/or nrETS) (Razafimandimbison et al., 2008, 2014; Antonelli et al., 2009; Rydin et al., 2009b; Razafimandimbison and Rydin, 2019). Such studies laid the foundation of the phylogenetic understanding within Rubioideae and the rest of the family. Recently, Rydin et al. (2017) and Wikström et al. (2020) used organellar genome scale datasets to reconstruct the phylogeny of the Rubiaceae family. Wikström et al. (2020) also analyzed nuclear ribosomal cistron data. Their results were mostly well supported and corroborated the overall picture of intertribal-relationships within Rubioideae, although high support values were not always achieved. Furthermore, results from the three different genomic compartments were not fully consistent (Rydin et al., 2017; Wikström et al., 2020). For example, deep-branching relationships within Rubioideae showed well supported yet conflicting tree topologies with either Ophiorrhizeae, a clade comprising Colletoecemateae and Urophylleae, or a clade comprising Colletoecemateae as sister to an Ophiorrhizeae + Urophylleae clade, resolved as sister group to the remaining subfamily. Another example of supported conflict was revealed by analysis of nuclear ribosomal data, which placed Coussareeae as sister to the Spermacoceae alliance, challenging the well documented sister-relationship between the Spermacoceae and Psychotrieae alliances in a number of previous studies (e.g., Bremer and Manen, 2000; Razafimandimbison et al., 2008; Rydin et al., 2009b). Relationships within the Psychotrieae and Spermacoceae alliances also differed between analyses of the different compartments, including deep splits within the Spermacoceae alliance, relationships among tribes of the Rubieae complex and the position of Gaertnereae in the Psychotrieae alliance. Antonelli et al. (2021) examined the

higher-level relationships in the entire Gentianales using target capture data, and while results were mostly consistent with those of previous studies, some surprising relationships were retrieved among their results. For instance, the sister relationship between Argostemmateae and the remaining tribes of the Spermacoceae alliance in their coalescent tree based on nuclear data, and the placement of Cyanoneuroneae nested within Psychotrieae alliance based on plastid data (Antonelli et al., 2021).

However, these family- and order-wide phylogenies have as a rule included only one representative taxon per sampled tribe and some key taxa have been unsampled. Furthermore, analysis of an organellar genome is generally considered to represent a single gene-tree within the species phylogeny (Gitzendanner et al., 2018; Doyle, 2022) and can thus fail to reflect the correct species tree due to processes such as incomplete lineage sorting (ILS) and hybridization (Nicholls et al., 2015; Wolf et al., 2018). Sampling a large number of presumably independently evolving genetic loci can avoid such problems and may even be necessary to infer the correct species tree (Degnan and Rosenberg, 2009; Nicholls et al., 2015; Ruane et al., 2015).

Targeted sequence capture uses short (often RNA) probes that are designed for the group of study to selectively capture target DNA regions from sequencing libraries and has emerged as a standard method for generating genome-scale nuclear multi-gene datasets for species tree inference in several plant groups (Johnson et al., 2019; Hale et al., 2020). The relative cost effectiveness and the fact that it works well also with degraded DNA, which is common among extractions of herbarium specimens, are some benefits of this approach (McKain et al., 2018; Johnson et al., 2019). The probe set used may be specifically designed for the group of study (e.g., Vatanparast et al., 2018; Sanderson et al., 2020) or designed to be universally applicable across larger groups such as the Angiosperms353 probe kit (Johnson et al., 2019). The large amount and heterogeneity of the data generated for phylogenomic studies do, however, not come without challenges. Factors such as poorly resolved gene trees due to low phylogenetic signal (Zhang et al., 2018), different types of data (Braun and Kimball, 2021), different data filtering strategies (Molloy and Warnow, 2018), and different underlying assumptions of phylogenetic inference methods such as concatenation- and coalescent-based methods (Roch and Steel, 2015) may all potentially affect accuracy of species tree inference.

Here, we attempt to resolve the phylogeny of the subfamily Rubioideae using large amounts of target capture data from the nuclear genome, and a much denser sampling of taxa, including several representatives of nearly all tribes of the subfamily, compared to previous work. We examine the impact of data type [coding sequence (CDS) and CDS + non-coding sequence], analytical approach (coalescence and concatenation), and potential paralogs (inclusion/exclusion of putative paralogous genes) on phylogenetic reconstruction. Our main aim is to improve the understanding of relationships within Rubioideae, mainly among tribes but also within tribes.

Materials and methods

Taxon sampling

One hundred and one Rubioideae species were selected to obtain a good representation of the subfamily. These species included representatives from all but two (the monogeneric tribes Foonchewieae and Aitchinsonieae) of the currently recognized tribes, and all but one non-monogeneric tribe was represented by more than one genus. For outgroup sampling we included twenty species to represent the major lineages of the remaining Rubiaceae, including representatives from the two other subfamilies and the two unplaced tribes Coptosapelteae and Luculieae. Three outgroup species from the Gentianales families Gentianaceae, Loganiaceae, and Apocynaceae were also selected. For 93 species, material was selected from vegetative tissue material (either silica dried material from field collections or from herbarium specimens) or from DNA aliquots already available from previous work. We also downloaded raw sequence data from the European Nucleotide Archive for 31 species available via the Plant and Fungal Tree of Life (PAFTOL) Research Program (Baker et al., 2022). Species and voucher information for all included taxa is provided in **Supplementary Table 1**.

Library preparation and target capture

DNA was extracted using a cetyl trimethylammonium bromide method (Doyle and Doyle, 1987). The plant tissue was pulverized using a TissueLyser LT (Qiagen, Hilden, Germany). Some samples were additionally cleaned with AMPure XP beads (Beckman Coulter, Indianapolis, IN, United States) or with a QIAquick polymerase chain reaction (PCR) kit (Qiagen, Hilden, Germany) according to the instructions provided by the manufacturers. DNA degradation was assessed by agarose gel (1%) electrophoresis and quantified on a Qubit 3 Flourometer (Thermo Fisher Scientific, Waltham, MA, United States) using the Qubit dsDNA HS kit. Samples with a large fraction of DNA fragments above 350 bp were placed in 96 microTUBE Plate wells and fragmented on a Covaris E220 Focused-ultrasonicator (Covaris, Woburn, MA, United States) using the program for a target insert size of 350 bp at Science for Life Laboratory (Solna, Sweden).

Libraries were prepared using a modified version of the Meyer and Kircher (2010) protocol. Briefly, the major steps of library preparation consisted of blunt-end repair, adapter ligation and adapter fill-in, followed by four separate index PCRs. End repair was performed in 40 μl reactions with 20 μl of DNA extract. AMPure bead cleanups after blunt-end repair and adapter ligation were performed using ratios of 0.9–1.8:1 AMPure to reaction volume. Adapter concentration in the ligation reaction was reduced to 0.25 μM of each adapter, and the cleanup step after adapter fill-in was substituted with

heat inactivation of the Bst polymerase at 80°C for 20 min following Kircher et al. (2012).

Each adapter-ligated library was then amplified with P5 and P7 dual-indexing primers in four separate PCR reactions to reduce amplification bias. One initial 12 cycle PCR per library was performed and the PCR products were loaded on a 1% agarose gel to verify amplification success and to determine an appropriate number of cycles for the remaining PCRs. Each 25 μl reaction contained 7 μl DNA library template and the following final concentrations: 1 × PCR Gold buffer, 2.5 mM MgCl2, 0.25 mM of each dNTP, 200 nM of each primer and 5 U AmpliTaq Gold. Reactions were subjected to the following thermocycling conditions: 94°C 12 min; 6–14 cycles of 94°C for 30 s, 60°C for 30 s, 72°C for 45 s; and a final extension of 72°C for 10 min. Individual PCR products for each sample were then pooled and cleaned using AMPure XP beads using ratios of 0.85-1:1 AMPure to reaction volume. The specific ratio used varied depending on DNA degradation, concentration and amount of unwanted short fragments (e.g., adapter-dimers) of the samples. The cleaned libraries were quantified using the Qubit dsDNA HS kit on a Qubit 3 Flourometer and fragment size distribution inspected with a high-sensitivity DNAchip on a Bioanalyzer 2100 (Agilent, Santa Clara, CA, United States).

Libraries of similar size were combined into 6-plex or 8-plex pools resulting in approximately equimolar 600 and 800 ng pools, respectively. Before pooling, apart from fragment size distribution, other factors, such as tissue source, number of PCR cycles during library preparation, age and library concentration were also considered. The pools were concentrated using either a miVac (Genevac, Ipswich, United Kingdom) or SpeedVac (Thermo Fisher Scientific, Waltham, MA, United States) at approximately 43°C. The pools were then enriched with the myBaits Expert Predesigned Panel (Arbor Biosciences, Ann Arbor, MI, United States) Angiosperms353 v1 (Catalog #308196; Johnson et al., 2019) following the manufacturer's protocol (v4).[1] Hybridization was carried out at 62°C for 24 or 36 h. Enriched products were amplified with KAPA HiFi (2×) HotStart ReadyMix PCR Kit (Roche, Basel, Switzerland) for 13–14 cycles with IS5_reamp. P5 and IS6_reamp.P7 primers (Meyer and Kircher, 2010) and subsequently cleaned using a 0.9:1 AMPure to reaction volume ratio. The hybridized and cleaned pools were quantified using the Qubit dsDNA HS kit and fragment size distribution inspected with a high-sensitivity DNAchip on a Bioanalyzer 2100. Finally, the enriched library pools were multiplexed at equimolar concentrations and sequenced on a NextSeq 500 using "Mid-Output" chemistry or NovaSeq 6000 using "NovaSeqXp" workflow in "S4" mode flowcell (Illumina, San Diego, CA, United States) with 151 bp paired-end reads at Science for Life Laboratory (Solna, Sweden).

1 http://www.arborbiosci.com/mybaits-manual

Data pre-processing

The Bcl to FastQ conversion was performed using bcl2fastq_v2.20.0.422 from the CASAVA software suite +, at Science for Life Laboratory (Solna, Sweden). The quality scale used was Sanger/phred33/Illumina 1.8. Further preprocessing of the obtained 151 bp paired-end reads was performed using utilities in the BBTools suite (BBTools, 2022). Dedupe or alternatively Clumpify was used to remove duplicate reads. BBduk was used to trim adapters, trim low-quality bases (Q < 20) and remove reads shorter than 36 bp. Dedupe and BBduk were used from within Geneious 11.1.5 (Kearse et al., 2012).

Gene assemblies

HybPiper v1.3.1 (Johnson et al., 2016) was used to assemble sequences for each gene. With the aim to increase gene recovery (gene length and number) the default target file for the Angiosperms353 kit was expanded by adding sequences of the Gentianales samples included in the mega353 target file produced by McLay et al. (2021) and the 348 sequences from the annotated *Coffea canephora* genome available via The Kew Tree of Life Explorer (Baker et al., 2022). The reads of library replicates from the same sample were combined before assembly. Read mapping was conducted using BWA v0.7.17 (Li and Durbin, 2009) and the coverage cut-off option was kept at the default value of eight for the SPAdes v3.15.2 (Bankevich et al., 2012) contig assembly. In addition to the default HybPiper coding sequence (CDS) output extracted with exonerate v2.2 (Slater and Birney, 2005) the optional HybPiper intronerate.py script was run to also extract so called supercontig sequences, which contain both CDS and non-coding flanking sequence. Recovery statistics were generated using the two HybPiper scripts get_seq_lengths.py and hybpiper_stats.py. The HybPiper script paralog_investigator.py was run to identify genes with paralog warnings. A HybPiper paralog warning is generated when HybPiper assembles multiple contigs covering more than 85% of the target length. In such a case HybPiper selects the sequence with highest sequencing coverage. If the copies have similar coverage, the copy with highest percent identity to the target sequence is chosen.

Alignment, dataset generation and phylogenetic analysis

The CDS and supercontig outputs for each target gene were aligned with MAFFT v7.467 (Katoh and Standley, 2013) with the L-INS-I algorithm and the additional –adjust direction flag. CDS alignments were aligned as amino acids and backtranslated using PAL2NAL v14 (Suyama et al., 2006). BMGE v1.12

(Criscuolo and Gribaldo, 2010) was used to trim sites with more than 90% gaps. The trimmed alignments were then concatenated using AMAS v1.0 (Borowiec, 2016), and Spruceup v2020.2.19 (Borowiec, 2019) was used to detect and trim outlier sequence windows from individual samples using the Jukes-Cantor-corrected distance method, a window size of 20 bp, an overlap size of 15 bp, a lognormal distribution and a cutoff value of 0.99. AMAS was then used to split the concatenated alignment into single-locus alignments and again trimmed with BMGE to remove sites with more than 90% gaps. The resulting alignments were used for phylogenetic inference. Alignment length, number and proportion of parsimony informative sites (PIS) and other alignment statistics were obtained using AMAS.

A total of four datasets were created. For each data type we created a dataset comprising the full set of genes (i.e., the direct HybPiper output), which we refer to as the full CDS dataset and full supercontig dataset. We also created a putative one-to-one ortholog dataset for each data type, which we refer to as the paralog-filtered CDS dataset and the paralog-filtered supercontig dataset. The two paralog-filtered datasets were created by conservatively removing any gene with at least one paralog warning from the respective full set of genes. The datasets were analyzed using a coalescent approach and a concatenation approach.

We used IQ-TREE 2 v2.0.3 (Minh et al., 2020) to infer a gene tree for each single gene alignment under the GTR + G model with support assessed with 1,000 ultrafast bootstrap replicates (Hoang et al., 2018). Following gene tree estimation, we collapsed nodes with less than 20% support using Newick Utilities v1.6 (Junier and Zdobnov, 2010) as this can help improve gene tree accuracy (Zhang et al., 2018). We then used the collapsed gene trees for species tree inference with a coalescent-based approach, using the quartet-based summary method ASTRAL III v5.7.8 (Zhang et al., 2018), which accounts for gene tree discordance due to ILS. Node support was assessed by local posterior probability (LPP; Sayyari and Mirarab, 2016). We also performed the polytomy test implemented in ASTRAL, which uses quartet gene tree frequencies to evaluate whether polytomies could be rejected at short branches (Sayyari and Mirarab, 2018). The normalized quartet score (NQS), which reflects the percentage of the gene tree quartets included in the species tree and part of the ASTRAL output, was used to assess the level of gene tree discordance for the respective datasets. To further examine gene tree discordance ASTRAL trees were annotated with quartet frequencies for alternative topologies using the –t 8 option in ASTRAL-III.

For each of the four datasets we also concatenated the single gene alignments to infer phylogenies in a concatenation framework. The concatenated matrices were analyzed using IQ-TREE 2 using a partitioned model (Chernomor et al., 2016), with each gene treated as a separate partition with a GTR + G model specified for each partition and allowing the possibility of separate rates among partitions. To assess branch support, ultrafast bootstrap supports (BS) were calculated based on 1,000 replicates.

Treeio (Wang et al., 2020) and ape (Paradis and Schliep, 2019) R packages (R Core Team, 2022) were used to plot the trees followed by editing in Inkscape v1.1.2 (Inkscape Project, 2022).

Results

Sequencing and assembly statistics

Sequencing and data filtration results can be found in **Supplementary Table 1**. Across all newly generated libraries the number of deduplicated and trimmed reads had a mean of 14,535,279. Across all libraries (i.e., including also the 31 PAFTOL samples downloaded from ENA) the number of deduplicated and trimmed reads had a mean of 11,653,388. The average library had 23% duplicate reads removed.

Assembly results are provided in **Supplementary Table 2**. At least a fraction of each of the 353 targeted genes were recovered in at least five taxa. Across the newly sequenced samples, the average sample had 336, 312, and 263 genes with sequences at least 25, 50, and 75% of the average target length, respectively, and a total gene length of 245,218 bp. Across all samples the average sample had 323, 291, and 237 genes with sequences at least 25, 50, and 75% of the average target length, respectively, and a total gene length of 228,644 bp. In addition to the targeted coding regions, large amounts of non-targeted sequence data were recovered. The average total length of recovered supercontig (coding sequence and non-coding flanking sequence) data was 710,450 and 661,303 bp for the newly sequenced samples and all samples, respectively. Across the full taxon sample, HybPiper gave paralog warnings for at least one sample in 167 of 353 genes. On average, samples had nine paralog warnings.

Dataset characteristics

The main characteristics of the four assembled datasets are summarized in **Table 1** and full statistics for each single locus alignment are provided in **Supplementary Table 3**. Across the 353 loci the average final alignment had a taxon coverage of 94% (117/124 species), and a length of 880 and 2,989 bp for the CDS and supercontig datasets, respectively. The total concatenated length of the full CDS dataset was 310,806 bp and the full supercontig dataset was 1,055,164 bp. The exclusion of the putatively paralogous genes (i.e., the genes flagged with paralog warnings by HybPiper) resulted in 186 alignments each for the paralog-filtered datasets with a total concatenated length of 181,088 and 632,932 bp for the CDS and supercontig datasets,

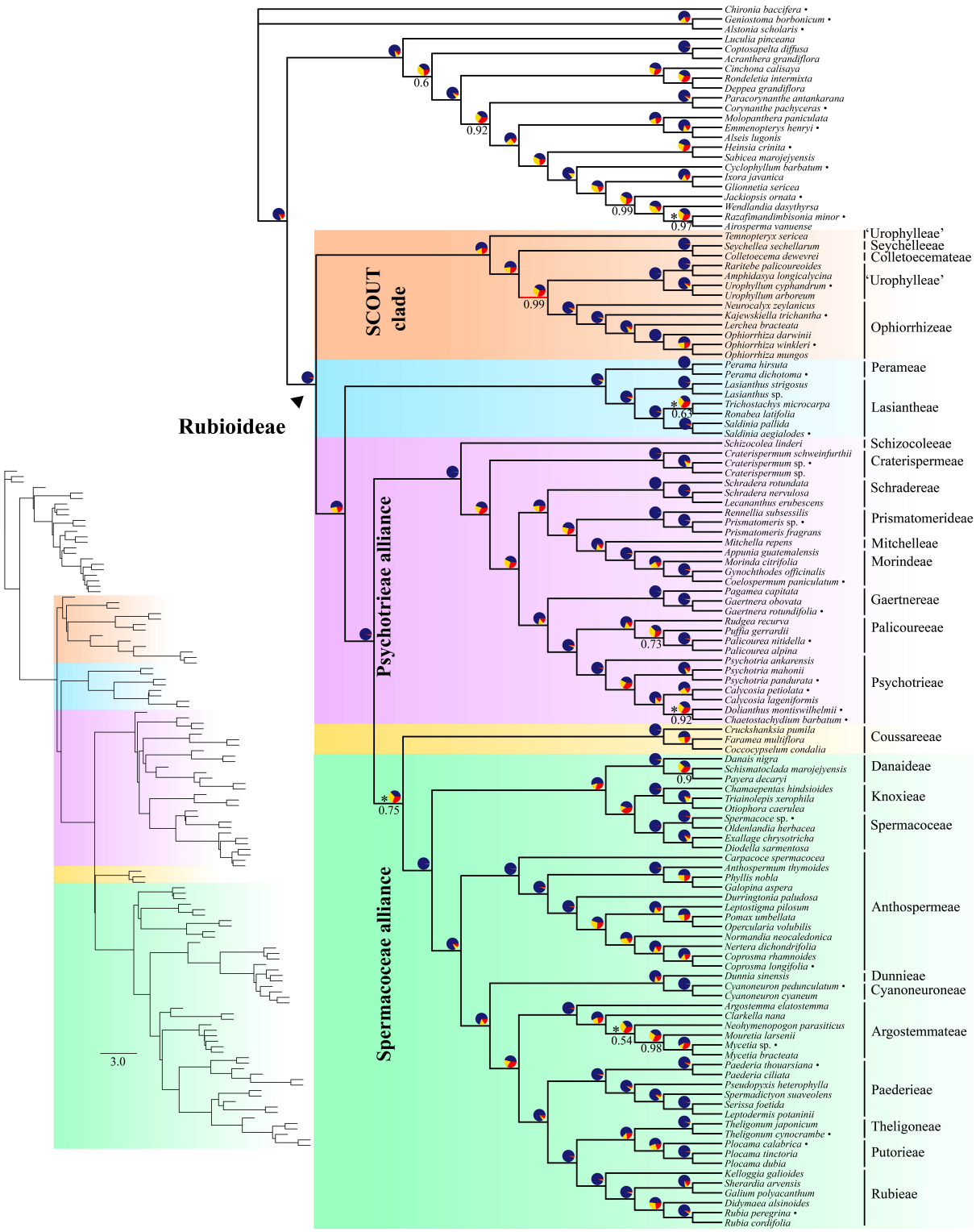


FIGURE 1
Coalescent-based species tree estimated using ASTRAL on the full supercontig dataset. Numbers below branches denote local posterior probability (LPP) support values. Only support values smaller than 100% are shown. Pie charts show relative frequencies of the three quartet topologies around the branch (blue = congruent with species tree, yellow = first alternative topology, red = second alternative topology). Asterisks next to pie charts indicate failure to reject the hypothesis that the branch is a polytomy. Bullets after species names indicate samples downloaded from ENA. Inset shows branch lengths in coalescent units.

10.3389/fpls.2022.967456

TABLE 1 Characteristics of assembled datasets used for phylogenetic inference.

Dataset	# Of loci	Concatenated length	# Of PIS (%)	Average taxon coverage (%)	Average alignment length	Average PIS per locus	Average percentage PIS per locus
Full supercontig	353	1,055,164	876,813 (83.1%)	117/124 (94.4%)	2,989	2,484	82.7
Full CDS	353	310,806	169,772 (54.6%)	117/124 (94.4%)	880	481	53.6
Paralog-filtered supercontig	186	632,932	526,877 (83.2%)	115/124 (92.7%)	3,403	2,833	82.9
Paralog-filtered CDS	186	181,088	99,394 (54.9%)	115/124 (92.7%)	974	534	53.7

PIS, parsimony informative sites.

TABLE 2 Phylogenetic inference performance of the assembled datasets for attributes under consideration.

	Phylogenetic inference approach					
	Coalescence (ASTRAL)				Concatenation (IQ-TREE)	
Dataset	Normalized quartet score	# Of branches below < 95% ingroup| global	# Of branches for which a polytomy could not be rejected. ingroup| global	Average LPP	# Of branches below < 95% ingroup| global	Average BS
Full supercontig	0.930	6| 8	4| 5	0.983	0| 0	99.9
Full CDS	0.880	13| 17	8| 11	0.964	7| 10	97.8
Paralog-filtered supercontig	0.939	8| 11	9| 10	0.973	1| 4	99.6
Paralog-filtered CDS	0.882	17| 23	14| 19	0.945	6| 10	97.5

respectively. On average, supercontig alignments contained over five times more PIS than CDS alignments.

Comparison of data types and inclusion/exclusion of potential paralogous genes

The performance of the four datasets on branch support, gene tree discordance (NQS values) and ability to reject polytomies are summarized in Table 2. Across both gene sets (i.e., inclusion/exclusion of putatively paralogous genes) and analytical approaches, the addition of non-coding sequences increased the average branch support, number of branches where a polytomy could be rejected, number of highly supported nodes, and gene tree concordance (i.e., higher NQS values). For the coalescence-based analyses of the full and paralog-filtered datasets there were nine (ingroup = seven) and 12 (ingroup = nine) more strongly supported nodes when using supercontigs instead of CDS alone, respectively. For the concatenated analyses of the full and paralog-filtered datasets there were 10 (ingroup = seven) and six (ingroup = five) more strongly supported nodes when using supercontigs instead of CDS alone, respectively. The number of branches where a polytomy could be rejected using the polytomy test in ASTRAL in the analyses of the full and paralog-filtered datasets was also higher when supercontigs were used instead of CDS alone, increasing with six (ingroup = four) and nine (ingroup = five) branches, respectively. Across both gene sets, supercontigs increased average BS support with 2.1% for the full and paralog-filtered datasets. Across both gene sets, supercontigs increased average LPP support with 1.9 and 2.8% for the full and paralog-filtered datasets, respectively. Across both gene sets the addition of flanking regions resulted in higher NQS values, increasing with 0.050 and 0.057 for the full and paralog-filtered datasets, respectively.

Across both data types and analytical approaches, the exclusion of genes with putative paralogs reduced the average branch support, number of branches where a polytomy could be rejected, and number of highly supported nodes, except for the concatenated analyses of CDS data where the exclusion of putatively paralogous genes resulted in one more well-supported ingroup branch. Excluding putatively paralogous genes from the supercontig data, the number of strongly supported nodes was reduced by four (ingroup = one) for the concatenation-based analysis. Excluding putatively paralogous genes from the supercontig and CDS data, the number of strongly supported nodes was reduced by three (ingroup = two) and six (ingroup = four) nodes for the coalescence-based analyses, respectively. Excluding putatively paralogous genes from supercontig and CDS data, the number of branches where a polytomy could be rejected decreased by five (ingroup = five) and eight (ingroup = six) branches, respectively. Across both data types, excluding putatively paralogous genes decreased average BS support by 0.3% for the full and paralog-filtered datasets. Across both gene sets, excluding putatively paralogous genes decreased average LPP support by 1 and 1.9% for the supercontig and CDS datasets, respectively. However, across

both data types the removal of putatively paralogous genes resulted in slightly higher NQS values, with an increase of 0.002 and 0.009 for the CDS and supercontig datasets, respectively.

Phylogenetic results

The inferred species tree topologies were highly similar regardless of method (coalescence- or concatenation-based), data type (CDS or supercontigs) and inclusion/exclusion of potentially paralogous genes (**Figures 1**, **2** and **Supplementary Figures 1–6**). The few topological conflicts were often not well supported (i.e., were supported by less than 95%). Overall, both the addition of flanking regions and inclusion of all genes increased statistical support and the power to reject polytomies. Therefore, we in the following, focus on the results obtained from the analyses of the full supercontig dataset (**Figures 1**, **2**).

Monophyly of Rubioideae, alliances, and tribes

Rubioideae, the Spermacoceae and Psychotrieae alliances, and all tribes except Urophylleae were highly supported as monophyletic (**Figures 1**, **2**). Urophylleae as delimited by Smedmark et al. (2008) was never monophyletic in any of the inferred species trees. However, Urophylleae excluding *Temnopteryx* was always highly supported as monophyletic (**Figures 1**, **2**), and this clade will hereafter be referred to as Urophylleae sensu stricto (s.s.).

Rubioideae backbone

Colletoecemateae, Ophiorrhizeae, Seychelleeae, Urophylleae s.s., and the genus *Temnopteryx* formed a clade (hereafter referred to as the SCOUT clade) sister to remaining Rubioideae, followed by a Lasiantheae + Perameae clade, the Psychotrieae alliance and a clade that joins the tribe Coussareeae and the Spermacoceae alliance (**Figures 1**, **2**). All these relationships were, with one exception, strongly supported and polytomies were rejected. The exception was the sister relationship between Coussareeae and the Spermacoceae alliance, which was strongly supported (BS = 100) in the concatenated analysis (**Figure 2**) but had low support (LPP = 0.75) in the coalescence-based tree (**Figure 1**) and a polytomy could not be rejected.

SCOUT clade

Relationships within the SCOUT clade differed between analytical approaches. The coalescence-based tree resolved the genus *Temnopteryx* as sister to the remaining members, followed by a Seychelleeae + Colletoecemateae clade, and a Urophylleae s.s. + Ophiorrhizeae clade (**Figure 1**). The concatenation-based tree instead resolved Ophiorrhizeae as sister to the Seychelleeae + Colletoecemateae clade (**Figure 2**). Support for these sets of relationships was high for all nodes and polytomies were rejected.

Psychotrieae alliance

The tribe Schizocoleeae and Craterispermeae were successive sisters to the remaining Psychotrieae alliance; this last clade was in turn resolved in two sister lineages: A clade formed by Schradereae, Prismatomerideae, Morindeae and Mitchelleae and a clade uniting Gaertnereae and Psychotrieae + Palicoureeae (**Figures 1**, **2**). Within the former clade, Schradereae and Prismatomerideae are successive sisters to Morindeae plus Mitchelleae (**Figures 1**, **2**). Support for this set of relationships was high for all nodes and polytomies were rejected.

Spermacoceae alliance

In the Spermacoceae alliance a clade that joins Danaideae and Spermacoceae + Knoxieae was together sister to the remaining tribes, followed by Anthospermeae, a clade that joins Dunnieae + Cyanoneuroneae, Argostemmateae, Paederieae, and a clade that joins Theligoneae and Putorieae and Rubieae (**Figures 1**, **2**). Support for this set of relationships was high for all nodes and polytomies were rejected.

Discussion

The most comprehensive multigene phylogenetic analysis of Rubioideae yet published is presented here. The vast majority of nodes was strongly supported (≥ 95%) in both the coalescence-based and concatenation-based phylogenies (**Figures 1**, **2**). We analyzed the data using a coalescent approach as well as a concatenation approach to phylogenetic inference of this group of plants, and we tested the inclusion/exclusion of putatively paralogous genes and the added information of non-targeted flanking regions in order to explore if relationships are reliant on a specific dataset or method. Leveraging substantial amounts of nuclear low-copy genetic data from a comprehensive taxon sample allowed us to infer a robust phylogenetic framework for the Rubioideae, potentially resolving and clarifying previously contentious relationships across the phylogeny of the group. For example, all inter-tribal relationships within the Spermacoceae and Psychotrieae alliances are robustly supported. Our study further supports the sister relationship between Coussareeae and the Spermacoceae alliance previously reported by Wikström et al. (2020) based on nuclear ribosomal cistron data. Within the Psychotrieae alliance, the Southeast Asian genus *Lecananthus* is nested in *Schradera,* more closely related to the Asian species *Schradera nervulosa* than either is to the neotropical *Schradera rotundata. Clarkella* is clearly included in the Argostemmateae, and *Pseudopyxis* in the Paederieae. The last three results are not unexpected considering morphological and geographic data. Furthermore, *Temnopteryx* is excluded from all currently described tribes of Rubiaceae; it is sister to remaining taxa in a well-supported SCOUT clade also comprising the tribes Seychelleeae, Colletoecemateae, Ophiorrhizeae and Urophylleae

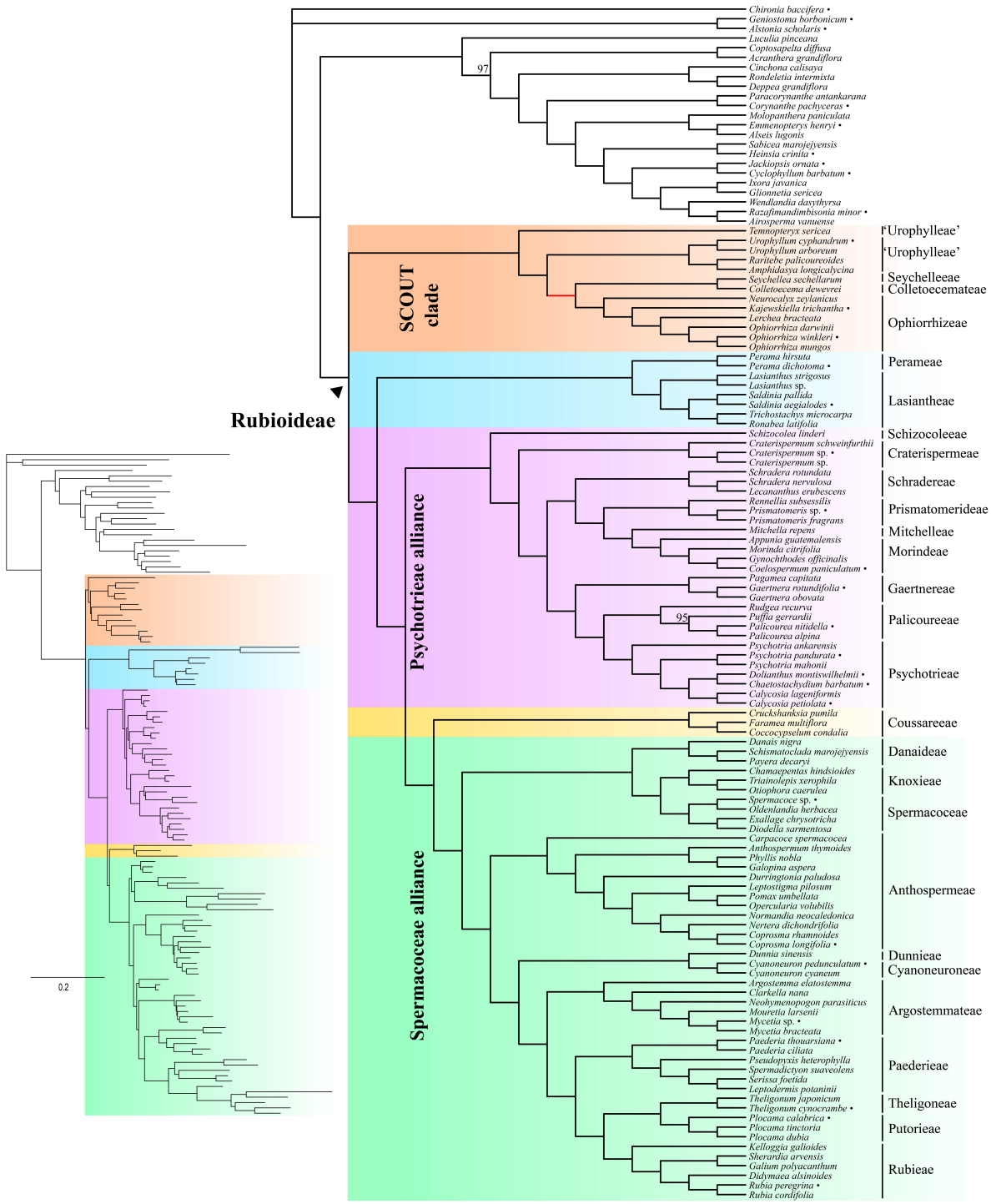


FIGURE 2
Concatenation-based tree estimated using IQ-TREE on the full supercontig dataset. Numbers above branches denote ultrafast bootstrap (BS) support values. Only support values smaller than 100% are shown. Bullets after species names indicate samples downloaded from ENA. Inset shows branch lengths in number of substitutions per site.

s.s., which together are supported as sister group to the remaining Rubioideae. Our study also shows that target capture data can resolve phylogenetic relationships with high confidence even in situations involving short branches, especially so when the combined information of coding and non-coding regions are used. Overall, our results indicate that ILS due to rapid

diversification is likely one of the major underlying causes responsible for most of the phylogenetic incongruences at short branches in the Rubioideae phylogeny.

Impact of potential paralogs, data type, and analytical method on phylogenetic inference

Inclusion of paralogous sequences can have important consequences for phylogenetic inference (Fitch, 1970; Yang and Smith, 2014). However, the topological results based on the full and paralog-filtered datasets mainly agree and statistical support increases when all genes are used. These factors suggest that (potential) paralogy did not change the topological results in any significant way, although the NQS values indicated slightly less gene tree discordance in the paralog-filtered data. This is in line with the results of Yan et al. (2021), which showed that ASTRAL and other coalescence-based methods are robust to species tree inference also in the presence of paralogs. Their study did, however, not include analyses of concatenated datasets, in which outlier genes have been shown to have extreme impact on topological results (Brown and Thomson, 2017). We used the target-capture data assembly HybPiper pipeline to assemble our datasets. This pipeline identifies paralogous copies and by default selects one copy based on sequencing coverage and percent identity to the target sequence. In other words, one copy per sample for each gene is selected and the approach is often applied to assemble target capture datasets (e.g., Antonelli et al., 2021; Clarkson et al., 2021; Maurin et al., 2021). However, this method may also flag genes with allelic variants rather than paralogs (Johnson et al., 2016) and may not uncover all paralogs (Zhou et al., 2022). Hence, both over- and underestimation of the number of detected putative paralogs is a possible outcome. Another common approach to deal with paralogs is to exclude entire genes that show evidence of paralogy, e.g., by removing putatively paralogous genes flagged by HybPiper (e.g., Larridon et al., 2020; Christe et al., 2021; Kuhnhäuser et al., 2021). Here, this approach resulted in a severe reduction of available sequence data left for species tree inference, which is common when many species are sampled (Emms and Kelly, 2018; Jones et al., 2019). This strict reliance on one-to-one orthologs led to an overall decrease in support and is likely to be an overly conservative approach in many phylogenetic contexts. Although (potential) paralogy did not seem to have any significant impact on the topological results presented in this paper, a more thorough analysis of paralogy may be worthwhile for future studies of subclades (e.g., genera) of Rubioideae. For example, identified paralogous copies could be used as additional loci (Gardner et al., 2021).

One advantage of targeted enrichment sequencing is that it facilitates assembly of non-targeted exon-flanking regions, including introns and sequence 5′ and 3′ to CDSs (Weitemier et al., 2014). Using the combined information of targeted CDS and non-targeted non-coding flanking sequence (supercontigs) improved overall statistical support as measured by number of highly supported nodes and average statistical support when compared to analyses of targeted CDS regions only. This finding is corroborated by other studies that have demonstrated increased statistical support for relationships by addition of flanking regions (e.g., Jones et al., 2019; Bagley et al., 2020; Gardner et al., 2021; Thomas et al., 2021). Addition of flanking regions also increased gene tree concordance and the power to reject polytomies with the polytomy test implemented in ASTRAL. Highly variable non-coding regions can be difficult to align but conserved flanking exons can help improve accuracy by anchoring the alignment (Gardner et al., 2021). Non-coding regions generally have higher evolutionary rates relative to CDS and should therefore contain more phylogenetic information, which may be necessary in order to resolve rapid speciation events (Chen et al., 2017). On the other hand, the higher variability (both in length and evolutionary rate) of non-coding regions may lead to higher degrees of noise. The overall higher statistical support we obtained using supercontig sequences and higher NQS values indicate that potential noise is overcome by the increased signal contained in these larger datasets.

It is notable, however, that there is one supported intertribal conflict between the paralog-filtered CDS and supercontig coalescence-based trees. While the analysis of the paralog-filtered supercontig data supported a Knoxieae + Danaideae clade (LPP = 0.96; **Supplementary Figure 3**), the paralog-filtered CDS data supported a Knoxieae + Spermacoceae clade (**Supplementary Figure 5**; LPP = 1). The latter relationship is highly supported in all other analyses in this study (including the concatenated analysis of the paralog-filtered dataset) and is also well established based on previous analyses of organellar and nuclear ribosomal DNA (Rydin et al., 2017; Wikström et al., 2020). Inspection of quartet frequencies shows that the two alternative quartet frequencies around the Knoxieae + Danaideae branch are not close (**Supplementary Figure 1**). This is contrary to the expectation of matching frequencies between the two alternative topologies if incongruence is due to ILS, indicating that sources of discordance other than ILS are involved, such as gene tree estimation error or gene flow (Degnan and Rosenberg, 2009; Leebens-Mack et al., 2019). The failure of the paralog-filtered dataset to resolve the Knoxieae + Spermacoceae relationship may be due to the much lower gene sampling in that dataset. However, the two alternative quartet frequencies around the Knoxieae + Spermacoceae branch in the full supercontig tree are also not close (**Figure 1**). Interestingly, the two alternative quartet frequencies around the Knoxieae + Spermacoceae branch in the two CDS trees (**Supplementary Figures 3, 5**) are similar and more indicative of ILS as the main source of discordance. A possible explanation for the patterns of quartet frequencies between analysis of CDS and supercontig data is

that the highly variable non-coding regions of the supercontigs introduce gene tree estimation error due to noise in this part of the tree. Another possible explanation could be that introgression of DNA is biased toward non-coding regions following hybridization.

Gene tree heterogeneity is widespread in multigene datasets (Edwards et al., 2016). Potential biological reasons for gene tree incongruence include ILS, hybridization, and gene duplication and loss (Maddison, 1997). Of these, ILS, which is modeled by the multispecies coalescent model (MSCM) (Pamilo and Nei, 1988), is the most prevalent and has so far received most attention (Edwards, 2009; Davidson et al., 2015). High levels of ILS are most likely to occur when there is a short time between speciation events, i.e., when internal branches of the species tree are short (Maddison, 1997; Whitfield and Lockhart, 2007). The concatenation approach combines the information from all available alignments into a single alignment and can mitigate low phylogenetic signal-to-noise problems (Philippe et al., 2005; de Queiroz and Gatesy, 2007). However, it ignores ILS and may, conversely to coalescence-based approaches, return highly supported but erroneous estimates of relationships in or near the anomaly zone, a region of tree space caused by successive rapid speciation events in the species tree, in which the most probable gene tree topology differs from the species tree topology (Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007; Liu and Edwards, 2009; Edwards et al., 2016; Mendes and Hahn, 2018).

Despite this drawback concatenation often performs well under many conditions, even in the presence of moderately high ILS levels (Bayzid and Warnow, 2013; Mirarab et al., 2016). Unlike concatenation, ASTRAL and several other coalescence-based methods can accommodate gene tree discordance due to ILS, and are statistically consistent under the MSCM (Mirarab and Warnow, 2015; Roch and Steel, 2015). Yet, coalescence-based approaches have been criticized for violating the MSCM assumptions such as error-free gene trees, absence of recombination within genes and free recombination between genes (Gatesy and Springer, 2014). While violations of the assumption of free recombination between loci can result in inaccurate phylogenetic estimates (Wang and Liu, 2016), both simulation and empirical studies have indicated that analyses using ASTRAL are largely robust to inclusion of recombinant loci (Lanier and Knowles, 2012; Wang and Liu, 2016; Folk et al., 2017; Morales-Briones et al., 2018). Nevertheless, coalescent-based methods can be sensitive to gene tree error, which can be alleviated using more informative genes and/or collapsing poorly supported relationships in gene trees prior to species tree inference (Zhang et al., 2018).

In Rubioideae, concatenation- and coalescence-based approaches generated highly similar topologies. However, one notable and highly supported topological conflict between the two approaches was detected: in the concatenated tree Ophiorrhizeae and Colletoecemateae + Seychelleeae formed a clade (BS = 100; **Figure 2**), whereas Ophiorrhizeae and Urophylleae s.s. formed a clade (LPP = 0.99; **Figure 1**) in the coalescence tree. This part of the tree has successive relatively short internal branches, a typical pattern of the anomaly zone, and indicate that the divergent placements of Ophiorrhizeae can be due to ILS and how it is differently accounted for in the two analytical approaches (Linkem et al., 2016). While inaccurate ortholog inference as well as gene tree error can generate gene tree incongruence, the pattern of gene tree quartet frequencies (**Figure 1**) with one main topology and balanced frequencies among the alternative topologies is more compatible with ILS as the main source of incongruence (Zou et al., 2008; Degnan and Rosenberg, 2009). It should be noted that the same incongruence is found also between the two analyses of the paralog-filtered dataset (**Supplementary Figures 3, 4**), but the support for the Ophiorrhizeae + Urophylleae s.s. branch was low (LPP = 0.91) and a polytomy could not be rejected. In contrast, this incongruence was not observed in the trees resulting from the analyses of the two CDS datasets, but except for the well-supported Colletoecemateae + Seychelleeae branch, all other intertribal relationships within the SCOUT clade were poorly supported in those trees and polytomies could not be rejected (**Supplementary Figures 1, 2, 5, 6**).

Phylogeny of Rubioideae

SCOUT clade

Studies addressing the deepest divergences in Rubioideae have often come to different conclusions but have most commonly involved the relative placements of the tropical African tribe Colletoecemateae, the Australasian tribe Ophiorrhizeae and the pantropical tribe Urophylleae. Analyses based on chloroplast sequence data have shown contradictory results; early studies based on Sanger sequencing of a few selected markers often found Colletoecemateae as sister to the remaining members of the subfamily (Robbrecht and Manen, 2006; Rydin et al., 2008, 2009a) but using a relaxed-clock model, Wikström et al. (2015) instead found Urophylleae as sister to the remaining tribes. More recent phylogenomic work has also resulted in topological incongruence; Ophiorrhizeae was sister to the remaining Rubioideae based on plastome data (Wikström et al., 2020), whereas the study by Rydin et al. (2017) based on mitochondrial data instead found Colletoecemateae + Ophiorrhizeae as sister to the remaining subfamily. The few analyses using nuclear data have consistently found a Colletoecemateae-Ophiorrhizeae-Urophylleae clade as the sister-group to the remaining Rubioideae, a result first reported by Rydin et al. (2009a) based on nrITS data and more recently also found based on nuclear ribosomal cistron (Wikström et al., 2020) and Angiosperm353 data (Antonelli et al., 2021). Based on plastid markers, the monogeneric

seychellean tribe Seychelleeae was recently found to be sister-taxon to the species-poor monogeneric tropical African tribe Colletoecemateae (Razafimandimbison et al., 2020), a relationship that is confirmed here. Our analyses consistently resolved a Seychelleeae-Colletoecemateae-Ophiorrhizeae-Urophylleae s.s. clade as the sister to the remaining subfamily Rubioideae, and we further show that the African genus *Temnopteryx* belongs in this clade, the SCOUT clade. Early classifications have differed in the tribal and subfamilial position of *Temnopteryx*, summarized by Khan et al. (2008), Smedmark et al. (2008). Khan et al. (2008) was the first phylogenetic study based on molecular to include *Temnopteryx,* and they showed that it belongs to Rubioideae although they did not resolve its position within the subfamily. In subsequent work based on molecular data (Smedmark et al., 2008, 2010; Smedmark and Bremer, 2011; Yang et al., 2016), *Temnopteryx* has been resolved as sister to the (remaining) tribe Urophylleae, although not always with high support. Here we instead find *Temnopteryx* strongly supported as sister to the remaining members of the SCOUT-clade (**Figures 1**, **2**).

Lasiantheae-Perameae

The second deepest split in the Rubioideae phylogeny separates a Lasiantheae-Perameae clade from the remaining members of the subfamily, i.e., a clade comprising the Psychotrieae and Spermacoceae alliances and the tribe Coussareeae. The sister-group relationship between the monogeneric tribe Perameae and Lasiantheae was first found by Andersson and Rova (1999) based on plastid *rps16* intron data and was considered surprising at the time, as there is no obvious morphological similarity between the two tribes. Although the tribe Perameae has not been as frequently sampled as Lasiantheae in molecular phylogenetic studies of Rubiaceae, the Lasiantheae-Perameae clade is well founded based on DNA sequence data with several subsequent studies supporting this relationship (e.g., Piesschaert et al., 2000; Smedmark et al., 2014; Antonelli et al., 2021; this study). While Perameae are tiny herbaceous plants with dry capsular fruits, Lasiantheae are woody, shrubby plants with fleshy drupes (Bremer and Manen, 2000; Smedmark et al., 2014). A feature they have in common is a solitary ovule in each locule, but the feature is found in several other members of Rubioideae as well (Bremer and Manen, 2000; Smedmark et al., 2014). The two tribes are thus morphologically distinct and we agree with previous authors (Andersson and Rova, 1999; Bremer and Manen, 2000) that merging the two tribes into Perameae should be avoided as it would create a morphologically undefinable taxon.

Coussareeae-Spermacoceae alliance

A notable result from the study by Wikström et al. (2020) was the placement of the tribe Coussareeae as sister to the Spermacoceae alliance on the basis of nuclear ribosomal data. The result conflicted with their own as well as previous results based on plastid data (Rydin et al., 2008, 2009a; Wikström et al., 2015, 2020; Neupane et al., 2017), plastid data + nrITS (Rydin et al., 2009b) and mitochondrial data (Rydin et al., 2017), which have all consistently supported the Coussareeae as sister to a clade comprised by the Spermacoceae and the Psychotrieae alliances. Our work (based on nuclear data) is congruent with the analyses of nuclear ribosomal cistron data by Wikström et al. (2020) regarding the relative positions of these three groups, but while the support for the sister relationship between the Coussareeae and the Spermacoceae alliance is high in the concatenated tree (BS = 100, **Figure 2**) it is relatively low in the coalescence-based tree (LPP = 0.75, **Figure 1**). The branch uniting Coussareeae and Spermacoceae alliance is short and gene tree heterogeneity high with quartet frequencies fairly even. Taken together, these findings indicate that ILS is the probable explanation for observed gene tree heterogeneity, and that a rapid speciation event may constitute the origin of these two sister clades.

Psychotrieae alliance

The nuclear phylogeny presented here includes representatives of all nine currently recognized tribes of the Psychotrieae alliance (Razafimandimbison et al., 2008, 2017) and shows, in contrast to previous studies based on Sanger-data, strong support across almost all relationships (including all inter-tribal relationships). Our study further supports the rare case of an evolutionary change from one-seeded carpels to many-seeded carpels found in the Psychotrieae alliance (Razafimandimbison et al., 2008), with Schradereae being the sole tribe with numerous ovules per locule. Our results are congruent with previously published phylogenies based on nuclear and mitochondrial data, although the studies by Rydin et al. (2017) and Wikström et al. (2020) did not include Schradereae and Antonelli et al. (2021) did not include Schradereae and Mitchelleae. Schradereae is here resolved as sister to the clade containing Prismatomerideae and the Morindeae-Mitchelleae clade. This clade is in turn sister to a clade comprising Gaertnereae and the Palicoureeae-Psychotrieae clade. The positions of the monogeneric African tribes Schizocoleeae and Craterispermeae as successive sisters to all other members of Psychotrieae alliance is consistent also with previous analyses based on plastid data. However, analyses of plastid data have found a sister-relationship between Gaertnereae and Prismatomerideae together sister to the Morindeae-Mitchelleae clade (Wikström et al., 2020; Antonelli et al., 2021), or Gaertnereae forming a clade with Schradereae, Morindeae and Mitchelleae together sister to the Palicoureeae-Psychotrieae clade with Prismatomerideae placed as sister to those two clades (Wikström et al., 2015).

Analyses based on combined plastid and nuclear ribosomal markers have largely produced results consistent with our results but have supported a Craterispermeae + Prismatomerideae clade (Razafimandimbison et al., 2008), or the

placement of Gaertnereae in a clade together with Schradereae, Prismatomerideae, Mitchelleae and Morindeae (Razafimandimbison et al., 2017). It is interesting to note that the nuclear results and mitochondrial results agree and are both in conflict with the plastid signal. Such discrepancies between results obtained with nuclear and mitochondrial data on one hand and plastid data on the other may be the result of old introgression events. However, the relatively short branch lengths and the quartet frequencies along the backbone nodes of the Psychotrieae alliance indicate relatively high levels of ILS during the early diversification of this clade.

Spermacoceae alliance

Resolving relationships in Spermacoceae alliance has been problematic, with relationships either unconvincingly supported or showing discordant topologies. In the Spermacoceae alliance our results support the position of the Danaideae-Knoxieae-Spermacoceae clade as sister taxon to the remaining members of the alliance. Several previous studies have shown results congruent with this, including a study based on mitochondrial data Rydin et al. (2017), the plastome-based phylogenomic analyses in Wikström et al. (2020), and analyses of a few selected plastid markers alone or in combination with nuclear ribosomal ITS (nrITS, e.g., Rydin et al., 2009a; Krüger, 2014; Wikström et al., 2015; Thureborn et al., 2019). Other analyses based on a few selected plastid markers, alone or in combination with nuclear ribosomal ITS, have not produced results congruent with ours, but have often found Danaideae as sole sister to the remaining members of the alliance (e.g., Bremer and Manen, 2000; Bremer and Eriksson, 2009; Rydin et al., 2009b; Yang et al., 2016). Analyses of the nuclear ribosomal cistron recovered yet another unexpected relationship with Anthospermeae sister to the Knoxieae-Spermacoceae clade, and Danaideae nested in a clade comprising the other sampled members of the alliance (Wikström et al., 2020). Further, the results presented by Antonelli et al. (2021) based on nuclear Angiosperms353 data showed surprisingly Argostemmateae (represented by one sample, *Mycetia* sp.) followed by Spermacoceae (represented by one sample, *Spermacoce* sp.) as successive sisters to the rest of Spermacoceae alliance. Those same samples were included in the present study (**Figures 1**, **2**), yielding other (more expected) topological placements of these samples. The discordance between our results and those of Antonelli et al. (2021) regarding the phylogenetic placement of these two samples may potentially be explained by the denser taxon sampling in the present study, for example in terms of tribes (11 vs. 8) and genera (40 vs. 10).

In the Rubieae complex, our results support the sister-relationship between Theligoneae and Putorieae and corroborate previous results based on nuclear ribosomal cistron Wikström et al. (2020) and Angiosperm353 data (Antonelli et al., 2021). Previous studies utilizing plastid data or a combination of plastid and nrITS data have either shown results consistent with our result (Yang et al., 2016; Antonelli et al., 2021; Rincón-Barrado et al., 2021) or have instead resolved Theligoneae and Rubieae as sister groups (e.g., Backlund et al., 2007; Bremer and Eriksson, 2009; Rydin et al., 2009b; Deng et al., 2017; Ehrendorfer et al., 2018; Wikström et al., 2020), a result also found when analyzing mitochondrial data (Rydin et al., 2017). While obvious morphological similarities supporting the Theligoneae + Rubieae clade seem to be lacking (Ehrendorfer et al., 2018) there are some morphological characters shared between some Putorieae species and members of clades within Rubieae (Natali et al., 1995; Ehrendorfer et al., 2018). Interestingly a recent study (Bordbar et al., 2021) found on the basis of the plastid *trnL-F* marker that *Plocama rosea* (Hemsl. ex Aitch.) M.Backlund and Thulin (= *Aitchisonia rosea* Hemsl. ex Aitch.) formed a clade with Rubieae, with Theligoneae and a clade containing the remaining sampled Putorieae/*Plocama* species as successive sisters to this clade. Based on those results the authors resurrected the monospecific genus *Aitchisonia* Hemsl. ex Aitch., and described the new monogeneric tribe Aitchisonieae to accommodate *A. rosea*. However, based on nrITS data the placement of *Plocama rosea* was inconclusive (Bordbar et al., 2021).

The sister group to the Rubieae complex is in our trees the tribe Paederieae, a relationship previously found in analyses based on nuclear and/or plastid data (Robbrecht and Manen, 2006; Rydin et al., 2009a,b; Wikström et al., 2015, 2020; Yang et al., 2016; Antonelli et al., 2021), although based on data from the mitochondrion this relationship was intervened by Argostemmateae (Rydin et al., 2017).

In our trees Anthospermeae, the Dunnieae + Cyanoneuroneae clade and Argostemmateae are supported as sequential sister groups to the Paederieae-Rubieae complex clade, a result fully congruent with the analyses of plastid data in Wikström et al. (2020). Other previous studies using plastid data and a combination of plastid and nuclear nrITS data have often been partly congruent with our results. The Anthospermeae-sister relationship has often been well supported but relationships among representatives of the remaining groups have generally been poorly supported (Rydin et al., 2009a,b; Wikström et al., 2015; Yang et al., 2016). Analyses of mitochondrial data have instead found Anthospermeae + Dunnieae, Paederieae and Argostemmateae as successive sisters to the Rubieae complex (Rydin et al., 2017). Analyses of the nuclear ribosomal cistron supported Anthospermeae as sister to the Knoxieae-Spermacoceae clade, and Danaideae nested in a clade containing the other sampled members of the alliance (Wikström et al., 2020). Previous analyses utilizing nuclear Angiosperm353 data (Antonelli et al., 2021) found Argostemmateae placed as sister to the remaining Spermacoceae alliance (represented by Spermacoceae, Cyanoneuroneae, Anthospermeae, Paederieae and the Rubieae complex) in their coalescence-based tree

(their concatenation-based tree was inconclusive except for the Paederieae-Rubieae complex phylogeny). However, our respective results are not fully comparable since Argostemmateae in our study includes also the single representative sample of Argostemmateae (*Mycetia* sp.) used in Antonelli et al. (2021) and the conflicting signal may thus be due to low sampling in their study relative to ours.

Our results support the close relationship between the two relatively recently described monogeneric tribes Dunnieae (China) (Rydin et al., 2009b) and Cyanoneuroneae (Borneo and Sulawesi) (Ginter et al., 2015). This result is congruent with Ginter et al. (2015) who, based on combined plastid and nuclear (nrETS and nrITS) data, found that those two tribes formed a clade that also included yet another recently described monogeneric tribe, the Foonchewieae from China (Wen and Wang, 2012). Thureborn et al. (2019) included representation from all these three tribes and found, based on plastid data, that they form a clade together with Argostemmateae (appendix B in Thureborn et al., 2019). Recent studies addressing major relationships in Rubiaceae have otherwise typically only included representation from one of these three tribes [for example, Antonelli et al. (2021) included Cyanoneuroneae and Rydin et al. (2017) and Wikström et al. (2020) included Dunnieae], but the close relationship between Foonchewieae and Dunnieae has been confirmed in several studies based on analyses of plastid data (Wikström et al., 2015; Yang et al., 2016). However, a highly unexpected placement of Cyanoneuroneae was found in the plastid tree of Antonelli et al. (2021); the Spermacoceae alliance excluding Cyanoneuroneae was strongly supported and Cyanoneuroneae was with strong support deeply nested in a clade comprising the sampled members of the Psychotrieae alliance. This result is not retrieved in other previous studies, nor in the results of the present study.

Infratribal relationships

Within tribes, our results reveal novel relationships and place a genus previously not included in phylogenetic analyses based on molecular data. Here we discuss intergeneric relationships within tribes whenever relevant and/or possible considering our sample of taxa.

Ophiorrhizeae

Within the Ophiorrhizeae, *Neurocalyx* is sister to the remaining tribe, and *Kajewskiella* is sister to *Lerchea* + *Ophiorrhiza* (**Figures 1**, **2**). The results are consistent with those of a recent study that investigated the phylogeny of Ophiorrhizeae using extensive species representation, five molecular markers and morphological considerations (Razafimandimbison and Rydin, 2019). Material for DNA-sequencing of *Kajewskiella* was unavailable to the authors at the time, but they predicted its inclusion in Ophiorrhizeae based on morphology, presumably sister to *Xanthophytum* (Razafimandimbison and Rydin, 2019). A later study included molecular data from *Kajewskiella* and confirmed its phylogenetic position in Ophiorrhizeae (Antonelli et al., 2021), although limited taxon sampling prevented further conclusions. The exact position of *Kajewskiella* within Ophiorrhizeae remains unresolved. The affinity to *Xanthophytum* was first suggested by Tange (1995) who discovered raphides in bract tissue in the inflorescences, "*. . .indistinguishable from those found in Xanthophytum*" (citation from Tange, 1995). The author found additional morphological indications of an affinity to *Xanthophytum* (Tange, 1995), and this was thus endorsed in the recent (greatly expanded) study of Ophiorrhizeae by some of us (Razafimandimbison and Rydin, 2019). Furthermore, Tange (1995) added information on *Kajewskiella* to Axelius's (1990) morphological data matrix of *Xanthophytum*, and reported that his parsimony analysis of the data placed *Kajewskiella* with *Xanthophytum papuanum, X. grandiporum, X. magnisepalum,* and *X. nitens*, a clade that had a derived position in Axelius's work (Axelius, 1990). There is thus ample morphological support for the reduction of *Kajewskiella* into *Xanthophytum*, as suggested by Tange (1995), but the hypothesis remains to be tested using molecular data from an adequate sample of species within the entire tribe, analyzed with state-of-the-art analytical tools.

Schradereae

In the tribe Schradereae, the Southeast Asian genus *Lecananthus* (Puff et al., 1998a) was recently shown to be nested in *Schradera* (Razafimandimbison et al., 2017), a result corroborated in the current study and further confirming the paraphyly of *Schradera* as delimited by Puff et al. (1998b). However, here *Lecananthus* is more closely related to the Asian species *Schradera nervulosa* than to the neotropical species *Schradera rotundata*.

Anthospermeae

We included representatives from 11 of the 12 genera of the Anthospermeae; only *Nenax* was not sampled since a recent study showed that species of *Nenax* are intermixed with those of *Anthospermum* in an *Anthospermum-Nenax* clade (Thureborn et al., 2019). Our results support the position of the South African genus *Carpacoce* as sister to a clade that unites an African clade and a Pacific clade, which is entirely congruent with results in Thureborn et al. (2019). Within the African clade, the positions of the southeastern Africa-centered genus *Galopina* and the Macaronesian genus *Phyllis* and their relationship(s) to *Anthospermum-Nenax* have been problematic with incongruent results and poor statistic support (Anderson et al., 2001; Yang et al., 2016; Thureborn et al., 2019). Here, *Galopina* and *Phyllis* form a highly supported clade (**Figures 1**, **2**), a relationship that has been suggested based on morphology (Sunding, 1979). It is worth noting that although this sister relationship is highly supported in all concatenated trees, only the supercontig dataset that

includes non-coding data had the power to reject the null hypothesis of a polytomy for this relatively short branch. The quartet frequencies (**Figure 1**) indicate that ILS contributes to a large proportion of the gene tree incongruence, which in combination with a relatively short branch suggest rapid speciation in the diversification history of this group. Within the Pacific clade, our results support the Australian genus *Durringtonia* as sister to the remaining clade, which in turn comprises (a) *Leptostigma* and *Pomax* + *Opercularia*, and (b) *Normandia* and *Coprosma* + *Nertera*. The analyses of nuclear data by Thureborn et al. (2019) placed *Durringtonia* in the latter clade but results were otherwise completely congruent with those presented here. Our results show that the subtribal classification of Anthospermeae, based mainly on flower and fruit characters (Puff, 1982), needs revision. The Australian subtribe Opercullariinae (*Pomax* and *Opercularia*) is monophyletic but is nested in the paraphyletic subtribe Coprosminae. Analyses of plastid data have previously indicated that both these subtribes are non-monophyletic (Thureborn et al., 2019) but support values were not significant. It should further be noted that Thureborn et al. (2019) detected some cases of supported cytonuclear discordance in the tribe. Generic interrelationships in Anthospermeae should be further investigated using genomic data.

Argostemmateae

Five of the genera we included in the present study were resolved in the Argostemmateae: *Argostemma, Clarkella, Neohymenopogon, Mouretia,* and *Mycetia. Argostemma* is sister to the remaining tribe. *Clarkella,* a small Asian herbaceous genus containing a single species (*Clarkella nana*), is here addressed for the first time using molecular data (but see Figure 2C in Yang et al., 2016), and the results show that it belongs in Argostemmateae, sister to *Neohymenopogon* + a *Mycetia–Mouretia* clade (**Figures 1**, **2**). *Clarkella* is currently placed in its own tribe Clarkelleae (Deb, 2001), but it was placed in Argostemmateae in earlier classifications (Verdcourt, 1958; Bremekamp, 1966). It was later excluded from Argostemmateae based on flower and pollen characters (Bremer, 1987) but both vegetative and fruit characters of *Clarkella* resemble those of some species of *Argostemma* (Puff and Chayamarit, 2008).

The intergeneric relationships within Argostemmateae are identical between the two inference methods we used (**Figures 1**, **2**) and all but one node (the *Neohymenopogon* + *Mycetia–Mouretia* clade in the coalescent tree, where a polytomy could not be rejected; **Figure 1**) are strongly supported. Our results differ, however, from those in previous studies (which are based on limited amount of molecular data, i.e., Rydin et al., 2009b; Ginter et al., 2015; Yang et al., 2016). Results in those studies are not always well supported and we too find indications of inconsistency regarding relationships in Argostemmateae. For example, in the analyses of the full CDS data, the coalescent tree supports a *Neohymenopogon* + *Mouretia* clade (**Supplementary Figure 1**), and the concatenation tree was inconclusive (i.e., support values were below 95%) for several relationships (**Supplementary Figure 2**) and inconsistent with the coalescent tree. Addition of data in the form of genes or longer sequences has been shown to lead to more congruence between species tree estimates (Cai et al., 2021; Gardner et al., 2021), and such a trend seems to be present also in Argostemmateae. Relationships in the tribe should nevertheless be investigated further, preferably also including the Asian and herbaceous genus *Leptomischus,* which recently was proposed to be sister to the remaining Argostemmateae based on plastid (*rbcL*) data (Razafimandimbison and Rydin, 2019).

Paederieae

We included five (*Leptodermis*, *Paederia*, *Pseudopyxis*, *Serissa,* and *Spermadictyon)* of the six currently recognized genera of Paederieae (Backlund et al., 2007; Rydin et al., 2009b). One of those is *Pseudopyxis* (*P. heterophylleae*), a genus here included in a molecular study for the first time. Its inclusion in the Paederieae is in line with Puff's (1982) classification of this tribe on the basis of morphology and geography. *Pseudopyxis* (three species) comprises perennial herbs occurring in China and Japan, and is here sister to a mainly woody Southeast Asian clade consisting of *Spermadictyon, Leptodermis,* and *Serissa.* Sister to those four genera is *Paederia,* a genus of pantropical and woody climbers. Our results agree well with the informal infratribal groupings suggested by Puff (1989) based on morphology and geography and are also consistent with previous molecular results based on plastid data (Backlund et al., 2007; Yang et al., 2016) as well as results based on a combination of plastid data and nrITS data (Rydin et al., 2009b). The Southeast Asian genus *Saprosma* is unfortunately not represented in our study. The genus was placed in Paederieae by Robbrecht (1993) based on morphology, and most subsequent work based on molecular data has since supported this, placing *Saprosma* either as sister to all other members of Paederieae (Rydin et al., 2009b) or sister to *Paederia* (Yang et al., 2016). It was however sister to the Rubieae complex in Backlund et al. (2007).

Data availability statement

The raw data generated for the present study are deposited in the European Nucleotide Archive (ENA) under study accession number PRJEB53647. The ENA sample accession numbers of all the samples are available in **Supplementary Table 1**. The target file used for HybPiper assembly and the assembled sequences are uploaded to Dryad Digital Repository, doi: 10.5061/dryad.d7wm37q44.

Author contributions

OT carried out the molecular experiments, post-sequencing bioinformatics analyses, and the phylogenetic analyses, and wrote the manuscript with input from all authors. All authors contributed to conception and design of the study, read, and approved the submitted version.

Funding

This project was funded by the Royal Swedish Academy of Sciences to CR.

Acknowledgments

We thank the herbaria AAU, BRI, CAS, CR, FUHM, GB, KLU, L, MEXU, MO, P, S, SBT, SEY, SPF, UPS, WAG, and WU for access to their collections and Anbar Khodabandeh, Bodil Cronholm, and Martin Irestedt for technical help related to library preparation and/or sequencing. We also thank the two reviewers for their constructive and helpful comments on the manuscript. We acknowledge support from the National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.967456/full#supplementary-material

References

Anderson, C. L., Rova, J. H. E., and Andersson, L. (2001). Molecular phylogeny of the tribe Anthospermeae (Rubiaceae): Systematic and biogeographic implications. *Aust. Syst. Bot.* 14:231. doi: 10.1071/SB00021

Andersson, L., and Rova, J. H. E. (1999). The *rps16* intron and the phylogeny of the Rubioideae (Rubiaceae). *Plant Syst. Evol.* 214, 161–186. doi: 10.1007/BF00985737

Antonelli, A., Clarkson, J. J., Kainulainen, K., Maurin, O., Brewer, G. E., Davis, A. P., et al. (2021). Settling a family feud: A high-level phylogenomic framework for the Gentianales based on 353 nuclear genes and partial plastomes. *Am. J. Bot.* 108, 1143–1165. doi: 10.1002/ajb2.1697

Antonelli, A., Nylander, J. A. A., Persson, C., and Sanmartín, I. (2009). Tracing the impact of the Andean uplift on Neotropical plant evolution. *Proc. Natl. Acad. Sci.* 106, 9749–9754. doi: 10.1073/pnas.0811421106

Axelius, B. (1990). The genus *Xanthophytum* (Rubiaceae). Taxonomy, phylogeny and biogeography. *Blumea Biodivers. Evol. Biogeogr. Plants* 34, 425–497.

Backlund, M., Bremer, B., and Thulin, M. (2007). Paraphyly of Paederieae, recognition of Putorieae and expansion of *Plocama* (Rubiaceae-Rubioideae). *Taxon* 56, 315–328. doi: 10.1002/tax.562006

Bagley, J. C., Uribe-Convers, S., Carlsen, M. M., and Muchhala, N. (2020). Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: Neotropical *Burmeistera* bellflowers as a case study. *Mol. Phylogenet. Evol.* 152:106769. doi: 10.1016/j.ympev.2020.106769

Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., et al. (2022). A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Syst. Biol.* 71, 301–319. doi: 10.1093/sysbio/syab035

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bayzid, M. S., and Warnow, T. (2013). Naive binning improves phylogenomic analyses. *Bioinformatics* 29, 2277–2284. doi: 10.1093/bioinformatics/btt394

BBTools (2022). *BBMap*. Available online at: https://sourceforge.net/projects/bbmap/. (accessed February 7, 2021).

Bordbar, F., Mirtadzadini, M., and Razafimandimbison, S. G. (2021). Phylogenetic re-assessment of the delimitation of *Plocama* and its species relationships and limits (Rubiaceae, Putorieae): Resurrection of the monospecific genus Aitchisonia and a description of trib. nov. Aitchisonieae. *Plant Syst. Evol.* 308:7. doi: 10.1007/s00606-021-01799-4

Borowiec, M. (2019). Spruceup: Fast and flexible identification, visualization, and removal of outliers from large multiple sequence alignments. *J. Open Source Softw.* 41635. doi: 10.21105/joss.01635

Borowiec, M. L. (2016). AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4:e1660. doi: 10.7717/peerj.1660

Braun, E. L., and Kimball, R. T. (2021). Data types and the phylogeny of Neoaves. *Birds* 2, 1–22. doi: 10.3390/birds2010001

Bremekamp, C. E. B. (1966). Remarks on the position, the delimitation and the subdivision of the Rubiaceae. *Acta Bot. Neerlandica* 15, 1–33. doi: 10.1111/j.1438-8677.1966.tb00207.x

Bremer, B. (1987). The sister group of the paleotropical tribe Argostemmateae: A redefined neotropical tribe Hamelieae (Rubiaceae, Rubioideae). *Cladistics* 3, 35–51. doi: 10.1111/j.1096-0031.1987.tb00495.x

Bremer, B., and Eriksson, T. (2009). Time tree of Rubiaceae: Phylogeny and dating the family, subfamilies, and tribes. *Int. J. Plant Sci.* 170, 766–793. doi: 10.1086/599077

Bremer, B., and Manen, J.-F. (2000). Phylogeny and classification of the subfamily Rubioideae (Rubiaceae). *Plant Syst. Evol.* 225, 43–72. doi: 10.1007/BF00985458

Brown, J. M., and Thomson, R. C. (2017). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Syst. Biol.* 66, 517–530. doi: 10.1093/sysbio/syw101

Cai, L., Xi, Z., Lemmon, E. M., Lemmon, A. R., Mast, A., Buddenhagen, C. E., et al. (2021). The perfect storm: Gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales. *Syst. Biol.* 70, 491–507. doi: 10.1093/sysbio/syaa083

Chen, M.-Y., Liang, D., and Zhang, P. (2017). Phylogenomic resolution of the phylogeny of laurasiatherian mammals: Exploring phylogenetic signals within coding and noncoding sequences. *Genome Biol. Evol.* 9, 1998–2012. doi: 10.1093/gbe/evx147

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* 65, 997–1008. doi: 10.1093/sysbio/syw037

Christe, C., Boluda, C. G., Koubínová, D., Gautier, L., and Naciri, Y. (2021). New genetic markers for Sapotaceae phylogenomics: More than 600 nuclear genes applicable from family to population levels. *Mol. Phylogenet. Evol.* 160:107123. doi: 10.1016/j.ympev.2021.107123

Clarkson, J. J., Zuntini, A. R., Maurin, O., Downie, S. R., Plunkett, G. M., Nicolas, A. N., et al. (2021). A higher-level nuclear phylogenomic study of the carrot family (Apiaceae). *Am. J. Bot.* 108, 1252–1269. doi: 10.1002/ajb2.1701

Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* 10:210. doi: 10.1186/1471-2148-10-210

Davidson, R., Vachaspati, P., Mirarab, S., and Warnow, T. (2015). Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC Genomics* 16:S1. doi: 10.1186/1471-2164-16-S10-S1

Davis, A. P., Govaerts, R., Bridson, D. M., Ruhsam, M., Moat, J., and Brummitt, N. A. (2009). A global assessment of distribution, diversity, endemism, and taxonomic effort in the Rubiaceae. *Ann. Mo. Bot. Gard.* 96, 68–78. doi: 10.3417/2006205

de Queiroz, A., and Gatesy, J. (2007). The supermatrix approach to systematics. *Trends Ecol. Evol.* 22, 34–41. doi: 10.1016/j.tree.2006.10.002

Deb, D. B. (2001). Study of floristics and plant taxonomy. *Phytotaxonomy* 1, 5–17.

Degnan, J. H., and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:e68. doi: 10.1371/journal.pgen.0020068

Degnan, J. H., and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340. doi: 10.1016/j.tree.2009.01.009

Deng, T., Zhang, J.-W., Meng, Y., Volis, S., Sun, H., and Nie, Z.-L. (2017). Role of the Qinghai-Tibetan Plateau uplift in the Northern Hemisphere disjunction: Evidence from two herbaceous genera of Rubiaceae. *Sci. Rep.* 7:13411. doi: 10.1038/s41598-017-13543-5

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Doyle, J. J. (2022). Defining coalescent genes: Theory meets practice in organelle phylogenomics. *Syst. Biol.* 71, 476–489.

Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1–19. doi: 10.1111/j.1558-5646.2008.00549.x

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., et al. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* 94, 447–462. doi: 10.1016/j.ympev.2015.10.027

Ehrendorfer, F., Barfuss, M. H. J., Manen, J.-F., and Schneeweiss, G. M. (2018). Phylogeny, character evolution and spatiotemporal diversification of the species-rich and world-wide distributed tribe Rubieae (Rubiaceae). *PLoS One* 13:e0207615. doi: 10.1371/journal.pone.0207615

Emms, D. M., and Kelly, S. (2018). STAG: Species tree inference from all genes. *bioRxiv* [preprint]. doi: 10.1101/267914

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113. doi: 10.2307/2412448

Folk, R. A., Mandel, J. R., and Freudenstein, J. V. (2017). Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Syst. Biol.* 66, 320–337. doi: 10.1093/sysbio/syw083

Gardner, E. M., Johnson, M. G., Pereira, J. T., Puad, A. S. A., Arifiani, D., Sahromi, et al. (2021). Paralogs and off-target sequences improve phylogenetic resolution in a densely sampled study of the breadfruit genus (*Artocarpus*, Moraceae). *Syst. Biol.* 70, 558–575. doi: 10.1093/sysbio/syaa073

Gatesy, J., and Springer, M. S. (2014). Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol. Phylogenet. Evol.* 80, 231–266. doi: 10.1016/j.ympev.2014.08.013

Ginter, A., Razafimandimbison, S. G., and Bremer, B. (2015). Phylogenetic affinities of *Myrioneuron* and *Cyanoneuron*, generic limits of the tribe Argostemmateae and description of a new Asian tribe, Cyanoneuroneae (Rubiaceae). *Taxon* 64, 286–298. doi: 10.12705/642.2

Gitzendanner, M. A., Soltis, P. S., Wong, G. K.-S., Ruhfel, B. R., and Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *Am. J. Bot.* 105, 291–301. doi: 10.1002/ajb2.1048

Hale, H., Gardner, E. M., Viruel, J., Pokorny, L., and Johnson, M. G. (2020). Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Appl. Plant Sci.* 8:e11337. doi: 10.1002/aps3.11337

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Inkscape Project (2022). *Inkscape. Draw freely*. Available online at: https://inkscape.org/ (accessed February 28, 2022).

Janssens, S. B., Groeninckx, I., De Block, P. J., Verstraete, B., Smets, E. F., and Dessein, S. (2016). Dispersing towards Madagascar: Biogeography and evolution of the madagascan endemics of the Spermacoceae tribe (Rubiaceae). *Mol. Phylogenet. Evol.* 95, 58–66. doi: 10.1016/j.ympev.2015.10.024

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., et al. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* 4:1600016. doi: 10.3732/apps.1600016

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., et al. (2019). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* 68, 594–606. doi: 10.1093/sysbio/syy086

Jones, K. E., Fér, T., Schmickl, R. E., Dikow, R. B., Funk, V. A., Herrando-Moraira, S., et al. (2019). An empirical assessment of a single family-wide hybrid capture locus set at multiple evolutionary timescales in Asteraceae. *Appl. Plant Sci.* 7:e11295. doi: 10.1002/aps3.11295

Junier, T., and Zdobnov, E. M. (2010). The Newick utilities: High-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* 26, 1669–1670. doi: 10.1093/bioinformatics/btq243

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Khan, S. A., Razafimandimbison, S. G., Bremer, B., and Liede-Schumann, S. (2008). Sabiceeae and Virectarieae (Rubiaceae, Ixoroideae): One or two tribes? New tribal and generic circumscriptions of Sabiceeae and biogeography of *Sabicea* s.l. *Taxon* 57, 7–23. doi: 10.2307/25065944

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40, e3–e3. doi: 10.1093/nar/gkr771

Krüger, Å (2014). *Systematics and biogeography of Western Indian Ocean region Rubiaceae: Examples from Danaideae, Hymenodictyeae, and Naucleeae*. [Doctoral dissertation]. Stockholm: Stockholm University.

Kubatko, L. S., and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56, 17–24. doi: 10.1080/10635150601146041

Kuhnhäuser, B. G., Bellot, S., Couvreur, T. L. P., Dransfield, J., Henderson, A., Schley, R., et al. (2021). A robust phylogenomic framework for the calamoid palms. *Mol. Phylogenet. Evol.* 157:107067. doi: 10.1016/j.ympev.2020.107067

Lanier, H. C., and Knowles, L. L. (2012). Is recombination a problem for species-tree analyses? *Syst. Biol.* 61, 691–701. doi: 10.1093/sysbio/syr128

Larridon, I., Villaverde, T., Zuntini, A. R., Pokorny, L., Brewer, G. E., Epitawalage, N., et al. (2020). Tackling rapid radiations with targeted sequencing. *Front. Plant Sci.* 10:1655. doi: 10.3389/fpls.2019.01655

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

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

Linkem, C. W., Minin, V. N., and Leaché, A. D. (2016). Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). *Syst. Biol.* 65, 465–477. doi: 10.1093/sysbio/syw001

Liu, L., and Edwards, S. V. (2009). Phylogenetic analysis in the anomaly zone. *Syst. Biol.* 58, 452–460. doi: 10.1093/sysbio/syp034

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.* 46, 523–536. doi: 10.1093/sysbio/46.3.523

Maurin, O., Anest, A., Bellot, S., Biffin, E., Brewer, G., Charles-Dominique, T., et al. (2021). A nuclear phylogenomic study of the angiosperm order Myrtales, exploring the potential and limitations of the universal Angiosperms353 probe set. *Am. J. Bot.* 108, 1087–1111. doi: 10.1002/ajb2.1699

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., and Yang, Y. (2018). Practical considerations for plant phylogenomics. *Appl. Plant Sci.* 6:e1038. doi: 10.1002/aps3.1038

McLay, T. G. B., Birch, J. L., Gunn, B. F., Ning, W., Tate, J. A., Nauheimer, L., et al. (2021). New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Appl. Plant Sci.* 9:as3.11420. doi: 10.1002/aps3.11420

Mendes, F. K., and Hahn, M. W. (2018). Why concatenation fails near the anomaly zone. *Syst. Biol.* 67, 158–169. doi: 10.1093/sysbio/syx063

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:db.rot5448. doi: 10.1101/pdb.prot5448

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534.

Mirarab, S., and Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, i44–i52. doi: 10.1093/bioinformatics/btv234

Mirarab, S., Bayzid, M. S., and Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst. Biol.* 65, 366–380. doi: 10.1093/sysbio/syu063

Molloy, E. K., and Warnow, T. (2018). To include or not to include: The impact of gene filtering on species tree estimation methods. *Syst. Biol.* 67, 285–303. doi: 10.1093/sysbio/syx077

Morales-Briones, D. F., Liston, A., and Tank, D. C. (2018). Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytol.* 218, 1668–1684. doi: 10.1111/nph.15099

Natali, A., Manen, J.-F., and Ehrendorfer, F. (1995). Phylogeny of the Rubiaceae-Rubioideae, in particular the tribe Rubieae: Evidence from a non-coding chloroplast DNA sequence. *Ann. Mo. Bot. Gard.* 82, 428–439. doi: 10.2307/2399892

Neupane, S., Lewis, P. O., Dessein, S., Shanks, H., Paudyal, S., and Lens, F. (2017). Evolution of woody life form on tropical mountains in the tribe Spermacoceae (Rubiaceae). *Am. J. Bot.* 104, 419–438. doi: 10.3732/ajb.1600248

Nicholls, J. A., Pennington, R. T., Koenen, E. J. M., Hughes, C. E., Hearn, J., Bunnefeld, L., et al. (2015). Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Front. Plant Sci* 6:710. doi: 10.3389/fpls.2015.00710

Pamilo, P., and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583. doi: 10.1093/oxfordjournals.molbev.a040517

Paradis, E., and Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Philippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. (2005). Phylogenomics. *Annu. Rev. Ecol. Evol. Syst.* 36, 541–562. doi: 10.1146/annurev.ecolsys.35.112202.130205

Piesschaert, F., Andersson, L., Jansen, S., Dessein, S., Robbrecht, E., and Smets, E. (2000). Searching for the taxonomic position of the African genus *Colletoecema* (Rubiaceae): Morphology and anatomy compared to an *rps16*-intron analysis of the Rubioideae. *Can. J. Bot.* 78, 288–304. doi: 10.1139/b00-002

Puff, C. (1982). The delimitation of the tribe Anthospermeae and its affinities to the Paederieae (Rubiaceae). *Bot. J. Linn. Soc.* 84, 355–377. doi: 10.1111/j.1095-8339.1982.tb00369.x

Puff, C. (1989). The affinities and relationships of the Japanese endemic *Pseudopyxis* (Rubiaceae-Paederieae). *Plant Species Biol.* 4, 145–155.

Puff, C., and Chayamarit, K. (2008). Additional to "Rubiaceae of Thailand. A pictorial guide to indigenous and cultivated genera. *Thai For. Bull. (Bot.)* 36, 70–80.

Puff, C., Buchner, R., and Greimler, J. (1998a). Revision of *Lecananthus* (Rubiaceae-Schradereae). *Blumea Biodivers. Evol. Biogeogr. Plants* 43, 337–346.

Puff, C., Greimler, J., and Buchner, R. (1998b). Revision of *Schradera* (Rubiaceae-Schradereae) in Malesia. *Blumea Biodivers. Evol. Biogeogr. Plants* 43, 287–335.

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Razafimandimbison, S. G., and Rydin, C. (2019). Molecular-based assessments of tribal and generic limits and relationships in Rubiaceae (Gentianales): Polyphyly of Pomazoteae and paraphyly of Ophiorrhizeae and *Ophiorrhiza*. *Taxon* 68, 72–91. doi: 10.1002/tax.12023

Razafimandimbison, S. G., Kainulainen, K., Senterre, B., Morel, C., and Rydin, C. (2020). Phylogenetic affinity of an enigmatic Rubiaceae from the Seychelles revealing a recent biogeographic link with Central Africa: Gen. nov. *Seychellea* and trib. nov. Seychelleeae. *Mol. Phylogenet. Evol.* 143:106685. doi: 10.1016/j.ympev.2019.106685

Razafimandimbison, S. G., Kainulainen, K., Wikström, N., and Bremer, B. (2017). Historical biogeography and phylogeny of the pantropical Psychotrieae alliance (Rubiaceae), with particular emphasis on the Western Indian Ocean Region. *Am. J. Bot.* 104, 1407–1423. doi: 10.3732/ajb.1700116

Razafimandimbison, S. G., Rydin, C., and Bremer, B. (2008). Evolution and trends in the Psychotrieae alliance (Rubiaceae)—A rarely reported evolutionary change of many-seeded carpels from one-seeded carpels. *Mol. Phylogenet. Evol.* 48, 207–223. doi: 10.1016/j.ympev.2008.03.034

Razafimandimbison, S. G., Taylor, C. M., Wikstrom, N., Pailler, T., Khodabandeh, A., and Bremer, B. (2014). Phylogeny and generic limits in the sister tribes Psychotrieae and Palicoureeae (Rubiaceae): Evolution of schizocarps in *Psychotria* and origins of bacterial leaf nodules of the Malagasy species. *Am. J. Bot.* 101, 1102–1126. doi: 10.3732/ajb.1400076

Rincón-Barrado, M., Olsson, S., Villaverde, T., Moncalvillo, B., Pokorny, L., Forrest, A., et al. (2021). Ecological and geological processes impacting speciation modes drive the formation of wide-range disjunctions within tribe Putorieae (Rubiaceae). *J. Syst. Evol.* 59, 915–934. doi: 10.1111/jse.12747

Robbrecht, E. (1993). Supplement to the 1988 outline of the classification of the Rubiaceae: Index to genera. *Opera Bot. Belg.* 6, 173–196.

Robbrecht, E. (1988). Tropical woody Rubiaceae. Characteristic features and progressions. Contribution to a new subfamilial classification. *Opera Bot. Belg.* 1, 251–267.

Robbrecht, E., and Manen, J.-F. (2006). The major evolutionary lineages of the coffee family (Rubiaceae, angiosperms). Combined analysis (nDNA and cpDNA) to infer the position of *Coptosapelta* and *Luculia*, and supertree construction based on *rbcL, rps16, trnL-trnF* and *atpB-rbcL* data. A new classification in two subfamilies, Cinchonoideae and Rubioideae. *Syst. Geogr. Plants* 76, 85–146. doi: 10.2307/20649700

Roch, S., and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.* 100, 56–62. doi: 10.1016/j.tpb.2014.12.005

Ruane, S., Raxworthy, C. J., Lemmon, A. R., Lemmon, E. M., and Burbrink, F. T. (2015). Comparing species tree estimation with large anchored phylogenomic and small Sanger-sequenced molecular datasets: An empirical study on Malagasy pseudoxyrhophiine snakes. *BMC Evol. Biol.* 15, 1–14. doi: 10.1186/s12862-015-0503-1

Rydin, C., Razafimandimbison, S. G., Khodabandeh, A., and Bremer, B. (2009b). Evolutionary relationships in the Spermacoceae alliance (Rubiaceae) using information from six molecular loci: Insights into systematic affinities of *Neohymenopogon* and *Mouretia*. *Taxon* 58, 793–810. doi: 10.1002/tax.583009

Rydin, C., Kainulainen, K., Razafimandimbison, S. G., Smedmark, J. E. E., and Bremer, B. (2009a). Deep divergences in the coffee family and the systematic position of *Acranthera*. *Plant Syst. Evol.* 278, 101–123. doi: 10.1007/s00606-008-0138-4

Rydin, C., Razafimandimbison, S. G., and Bremer, B. (2008). Rare and enigmatic genera (*Dunnia, Schizocolea, Colletoecema*), sisters to species-rich clades: Phylogeny and aspects of conservation biology in the coffee family. *Mol. Phylogenet. Evol.* 48, 74–83. doi: 10.1016/j.ympev.2008.04.006

Rydin, C., Wikström, N., and Bremer, B. (2017). Conflicting results from mitochondrial genomic data challenge current views of Rubiaceae phylogeny. *Am. J. Bot.* 104, 1522–1532. doi: 10.3732/ajb.1700255

Sanderson, B. J., DiFazio, S. P., Cronk, Q. C. B., Ma, T., and Olson, M. S. (2020). A targeted sequence capture array for phylogenetics and population genomics in the Salicaceae. *Appl. Plant Sci.* 8:e11394. doi: 10.1002/aps3.11394

Sayyari, E., and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* 33, 1654–1668. doi: 10.1093/molbev/msw079

Sayyari, E., and Mirarab, S. (2018). Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes* 9, 132. doi: 10.3390/genes9030132

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

Smedmark, J. E. E., and Bremer, B. (2011). Molecular systematics and incongruent gene trees of Urophylleae (Rubiaceae). *Taxon* 60, 1397–1406. doi: 10.1002/tax.605015

Smedmark, J. E. E., Eriksson, T., and Bremer, B. (2010). Divergence time uncertainty and historical biogeography reconstruction – an example from Urophylleae (Rubiaceae). *J. Biogeogr.* 37, 2260–2274. doi: 10.1111/j.1365-2699.2010.02366.x

Smedmark, J. E. E., Razafimandimbison, S. G., Wikström, N., and Bremer, B. (2014). Inferring geographic range evolution of a pantropical tribe in the coffee family (Lasiantheae, Rubiaceae) in the face of topological uncertainty. *Mol. Phylogenet. Evol.* 70, 182–194. doi: 10.1016/j.ympev.2013.09.007

Smedmark, J. E. E., Rydin, C., Razafimandimbison, S. G., Khan, S. A., Liede-Schumann, S., and Bremer, B. (2008). A phylogeny of Urophylleae (Rubiaceae) based on *rps16* intron data. *Taxon* 57, 24–32. doi: 10.2307/25065945

Sunding, P. (1979). "Origins of the macaronesian Flora," in *Plants and islands*, ed. D. Bramwell (London: Academic Press), 13–40.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Tange, C. (1995). The identity of *Siderobombyx* and a new species of *Xanthophytum* (Rubiaceae). *Nord. J. Bot.* 15, 575–581.

Thomas, A. E., Igea, J., Meudt, H. M., Albach, D. C., Lee, W. G., and Tanentzap, A. J. (2021). Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand *Veronica*. *Am. J. Bot.* 108, 1289–1306. doi: 10.1002/ajb2.1678

Thureborn, O., Razafimandimbison, S. G., Wikström, N., Khodabandeh, A., and Rydin, C. (2019). Phylogeny of Anthospermeae of the coffee family inferred using clock and nonclock models. *Int. J. Plant Sci.* 180, 386–402. doi: 10.1086/703353

Vatanparast, M., Powell, A., Doyle, J. J., and Egan, A. N. (2018). Targeting legume loci: A comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. *Appl. Plant Sci.* 6, e1036. doi: 10.1002/aps3.1036

Verdcourt, B. (1958). Remarks on the classification of the Rubiaceae. *Bull. Jard. Bot. LÉtat Brux.* 28, 209–290. doi: 10.2307/3667090

Wang, L.-G., Lam, T. T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., et al. (2020). Treeio: An R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* 37, 599–603. doi: 10.1093/molbev/msz240

Wang, Z., and Liu, K. J. (2016). A performance study of the impact of recombination on species tree analysis. *BMC Genomics* 17:785. doi: 10.1186/s12864-016-3104-5

Weitemier, K., Straub, S. C. K., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., et al. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Appl. Plant Sci.* 2:1400042. doi: 10.3732/apps.1400042

Wen, H.-Z., and Wang, R.-J. (2012). *Foonchewia guangdongensis* gen. et sp. nov. (Rubioideae: Rubiaceae) and its systematic position inferred from chloroplast sequences and morphology. *J. Syst. Evol.* 50, 467–476. doi: 10.1111/j.1759-6831.2012.00196.x

Whitfield, J. B., and Lockhart, P. J. (2007). Deciphering ancient rapid radiations. *Trends Ecol. Evol.* 22, 258–265. doi: 10.1016/j.tree.2007.01.012

Wikström, N., Bremer, B., and Rydin, C. (2020). Conflicting phylogenetic signals in genomic data of the coffee family (Rubiaceae). *J. Syst. Evol.* 58, 440–460. doi: 10.1111/jse.12566

Wikström, N., Kainulainen, K., Razafimandimbison, S. G., Smedmark, J. E. E., and Bremer, B. (2015). A revised time tree of the asterids: Establishing a temporal framework for evolutionary studies of the coffee family (Rubiaceae). *PLoS One* 10:e0126690. doi: 10.1371/journal.pone.0126690

Wolf, P. G., Robison, T. A., Johnson, M. G., Sundue, M. A., Testo, W. L., and Rothfels, C. J. (2018). Target sequence capture of nuclear-encoded genes for phylogenetic analysis in ferns. *Appl. Plant Sci.* 6, e01148. doi: 10.1002/aps3.1148

Yan, Z., Smith, M. L., Du, P., Hahn, M. W., and Nakhleh, L. (2021). Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Syst. Biol* 71, 367–381. doi: 10.1093/sysbio/syab056

Yang, L.-L., Li, H.-L., Wei, L., Yang, T., Kuang, D.-Y., Li, M.-H., et al. (2016). A supermatrix approach provides a comprehensive genus-level phylogeny for Gentianales: Phylogeny of Gentianales. *J. Syst. Evol.* 54, 400–415. doi: 10.1111/jse.12192

Yang, Y., and Smith, S. A. (2014). Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* 31, 3081–3092. doi: 10.1093/molbev/msu245

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

Zhou, W., Soghigian, J., and Xiang, Q.-Y. (2022). A new pipeline for removing paralogs in target enrichment data. *Syst. Biol* 71, 410–425. doi: 10.1093/sysbio/syab044

Zou, X.-H., Zhang, F.-M., Zhang, J.-G., Zang, L.-L., Tang, L., Wang, J., et al. (2008). Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol.* 9, R49. doi: 10.1186/gb-2008-9-3-r49

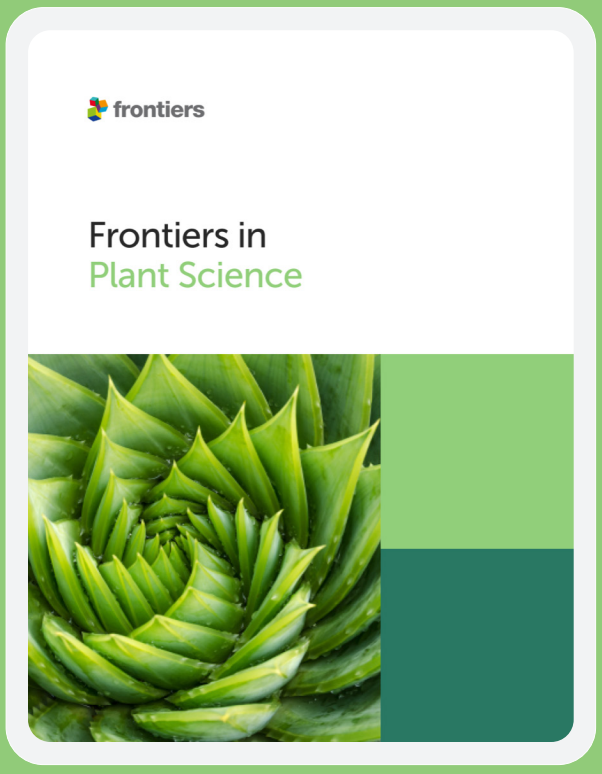
Frontiers in
Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact