PHYLOGENOMIC APPROACHES TO DEAL WITH PARTICULARLY CHALLENGING PLANT LINEAGES

EDITED BY: Marcial Escudero, Gonzalo Nieto Feliner, Lisa Pokorny, Daniel Spalink and Juan Viruel

PUBLISHED IN: Frontiers in Plant Science

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-367-5
DOI 10.3389/978-2-88966-367-5

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: researchtopics@frontiersin.org

PHYLOGENOMIC APPROACHES TO DEAL WITH PARTICULARLY CHALLENGING PLANT LINEAGES

Topic Editors:
Marcial Escudero, Sevilla University, Spain
Gonzalo Nieto Feliner, Real Jardín Botánico (RJB, CSIC), Spain
Lisa Pokorny, National Institute of Agricultural and Food Research and Technology, Spain
Daniel Spalink, Texas A&M University, United States
Juan Viruel, Royal Botanic Gardens, Kew, United Kingdom

Citation: Escudero, M., Feliner, G. N., Pokorny, L., Spalink, D., Viruel, J., eds. (2021). Phylogenomic Approaches to Deal With Particularly Challenging Plant Lineages. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-367-5

Table of Contents

frontiers in Plant Science

EDITORIAL
published: 24 November 2020
doi: 10.3389/fpls.2020.591762

Editorial: Phylogenomic Approaches to Deal With Particularly Challenging Plant Lineages

Marcial Escudero[1*†], *Gonzalo Nieto Feliner*[2†], *Lisa Pokorny*[2,3,4†], *Daniel Spalink*[5†] *and Juan Viruel*[4†]

[1] Department of Plant Biology and Ecology, University of Seville, Seville, Spain, [2] Department of Biodiversity and Conservation, Real Jardín Botánico, CSIC, Madrid, Spain, [3] Centre for Plant Biotechnology and Genomics (CBGP), Technical University of Madrid (UPM)-National Institute of Agriculture and Food Research and Technology (INIA), Madrid, Spain, [4] Royal Botanic Gardens, Kew, Richmond, United Kingdom, [5] Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, United States

Keywords: genotyping-by-sequencing, genome skimming, high-throughput sequencing, phylogenomics, RAD-seq, target capture, Hyb-Seq

Edited by:
Kathleen Pryer,
Duke University, United States

Reviewed by:
Donovan Bailey,
New Mexico State University, United States

****Correspondence:***
Marcial Escudero
amesclir@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: *05 August 2020*
Accepted: *02 November 2020*
Published: *24 November 2020*

Citation:
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

Editorial on the Research Topic

Phylogenomic Approaches to Deal With Particularly Challenging Plant Lineages

Molecular phylogenetics has been revolutionized in the past two decades by the development of increasingly cheap high-throughput sequencing (HTS) technologies. The transition from Sanger to HTS has simultaneously necessitated concerted efforts to develop molecular and computational capacities and technologies, as well as analytical tools to efficiently but rigorously interrogate the resulting massive datasets. Collectively, these approaches are often called "phylogenomics." Here, we use the term phylogenomics to refer to analyses of HTS data representing large portions of genomes. Indeed, phylogenomics has yielded unprecedented resolution and support to our understanding of the tree of life. At the same time, it has fully corroborated the early finding that the history of individual genes does not mirror the history of lineages and that rapid radiations, both recent and old, are difficult to resolve. The biological processes underlying gene tree discordance remain a challenge to detect and accommodate in most phylogenomic models, and particularly difficult lineages are not immediately clarified by the addition of orders of magnitude more data. Thus, while we have made significant progress in recent years, the remaining road to a resolved tree of life will continue to be challenging.

In this Special Issue, we focused on the use of phylogenomics to shine light into particularly challenging plant lineages. The papers presented herein are intended to illustrate how the application of new analytical approaches and laboratory protocols may tackle the processes underlying biological diversification, thus improving our understanding of lineage relationships through time. From establishing best practices in the development and use of targeted-enrichment bait kits, to formalizing parameter-exploring assembly pipelines, we expect this collection to serve as a useful resource in the planning and execution of successful phylogenomic projects.

Major challenges in phylogenomic approaches ultimately stem from the dynamism of genomes over evolutionary time, but this can be aggravated whenever lineages include hybridization and/or polyploidy in their history (Twyford and Ennos, 2012), since these processes can involve mixing of different phylogenetic signals and an extra amount of change triggered by the so-called genomic shock (McClintock, 1984). Although polyploidy is known to be pervasive in the history of vascular plants (Soltis and Soltis, 2012), detecting whole genome duplication (WGD) events is not trivial

due to the suite of genome readjustments that follow every single WGD event and continue over time until a new one occurs (Wendel, 2015). Stai et al. developed a pipeline to identify genome duplications in 14,709 gene families based on evolutionary, phylogenomic, and synteny analyses for the legume family. The study focused on the genus *Cercis* in the legume subfamily Cercidoideae, which they infer as sister to the other legume subfamilies (though see Koenen et al., 2020). They found evidence for a genome duplication (allotetraploidy) in the Cercidoideae subfamily and a set of independent genome duplications in the other legume subfamilies. Due to apomixis, hybridization and polyploidy, *Rubus* (Rosaceae) epitomizes the difficulties in unraveling relationships. Yet, using a target capture dataset comprising plastome and more than nine hundred nuclear loci, from a representative sampling of the rampant variation in ploidy in this genus, Carter et al. managed to untangle the complex phylogenetic relationships in the genus. This included inferring multiple hybridization events among the brambles as well as the biogeographic history of this genus, which involved a North American most recent common ancestor.

The Hawaiian species of the genus *Melicope* (Rutaceae) represent one of the major adaptive radiations of the Hawaiian Islands. Although reduced representation approaches are typically applied to microevolutionary questions, Paetzold et al. relied on RAD-seq to infer phylogenetic relationships, which were poorly resolved using Sanger sequencing. Their results drastically improved resolution of relationships within Hawaiian *Melicope* and, using ABBA-BABA tests, provided evidence for both ancestral and current hybridization events. Donkpegan et al. used a related technique, genotyping-by-sequencing (GBS), to infer diversification dates in African species of the tree genus *Afzelia* (Fabaceae), in which species delimitation and phylogenetic relationships among diploids and tetraploids remained unresolved. Their results suggest that a single biome shift took place from the savannah species, which are diploid, to the rainforest species, which are tetraploid. The implied pattern of an earlier Miocene diversification of the tropical savannah clade compared to the Pliocene diversification of the rainforest clade, is opposite to the one usually found for other groups in the region.

Cost-effective approaches such as genome skimming, which retrieves high-copy number nuclear (e.g., transposons and other repeats) and organellar (i.e., plastome, mitome) regions, have less power for shining light into complex evolutionary scenarios. However, using this technique in *Silene* section *Psammophilae* (Caryophyllaceae), del Valle et al. unveiled clear incongruence between morphology-based taxonomic boundaries and phylogeographic patterns inferred from whole plastomes. This result suggests a history of interspecific hybridization among Iberian populations of the five species that integrate this section, to the exclusion of the Balearic populations of only one (*S. cambessedesii)*. The same technique was used successfully by Moreno-Aguilar et al. to conclude that two enigmatic grass genera, *Megalachne* and *Podophorus* endemic to the Juan Fernandez Pacific archipelago, form a monophyletic group. They also inferred that a long-distance dispersal event gave rise to this group, from South American fescue populations, in the Miocene-Pliocene transition. It is also noticeable that this study is an example of museomics and, remarkably, sampling included a 164-year old type specimen of *Podophorus bromoides*, a species currently considered extinct.

Another challenge to all HTS methods is that handling massive amounts of genomic data generates uncertainty at different stages. Using the Mediterranean genus *Helianthemum* (Cistaceae), Martín-Hernanz et al. established a pipeline to explore the impact of different parameter settings during RAD-seq assembly on genotyping error rates. They found that different parameter configurations produced topologically congruent phylogenies, but also that minimizing error rates results in more reliable branch lengths which affected the accuracy of downstream analyses (i.e., divergence times and diversification rates).

Target sequencing capture approaches are gradually becoming predominant for tackling macroevolutionary questions in non-model organisms. One debated and yet unresolved question concerning these approaches is the selection of loci. It is assumed that specific bait kits designed for the group in question maximize capture. For instance, from a palm-specific enrichment panel targeting 4,184 genomic regions, Loiseau et al. selected 795 phylogenetically informative nuclear markers (PhyloPalm kit) to resolve relationships in palms (Arecaceae). They focused on a widely distributed group of neotropical palms—tribe Geonomateae—and obtained strongly supported topology for this group, whose relationships were previously far from settled. In another study focused on one of the largest tribes in the Asteraceae, Vernonieae, Siniscalchi et al. recovered c. 700 nuclear markers from a kit developed specifically for the family (Mandel et al., 2014) using Hyb-Seq (Weitemier et al., 2014), a HTS approach which combines target enrichment and genome skimming. Although sampling a small percentage of the 1,500 species in the tribe, the authors obtained complete resolution and high support in the phylogeny, substantially improving those in previous studies.

A specific bait kit was also designed for unraveling the neotropical radiation of the spiral ginger (*Costus*, Costaceae) by Valderrama et al. using available genomic resources for *Costus*. They obtained and used 832 loci for phylogenomic analyses—using both concatenation and coalescent-based species trees methods—with which the authors achieved a robust estimation of relationships despite high levels of gene tree conflict. By contrast, some studies rely on phylogenetically broad bait kits, such as the Angiosperms353 kit, which has been carefully designed to capture 353 single-copy nuclear loci across angiosperms, providing useful phylogenetic signal at different phylogenetic depth levels (Johnson et al., 2018). For instance, research in yet another tropical group in this issue uses this Angiosperm353 bait kit and the aforementioned Hyb-Seq approach to investigate the evolutionary history of the Papuasian *Schefflera* (Araliaceae) radiation (Shee et al.). By resolving both deep and shallow phylogenetic relationships, the authors show the efficacy of this universal bait kit, even when sampling herbarium material (including type specimens). They

also inferred a sequence of colonization events to explain the present-day distribution of this genus in Papuasia. Concerning the hot question of which bait kit to use in target enrichment approaches, Larridon et al. present an interesting comparison between the Angiosperm353 bait kit and a Cyperaceae-specific kit to unravel the rapid radiation of the C4 *Cyperus* clade. The results are as unexpected as fascinating.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by grants from the Spanish MICINN (PGC2018-17099608-B-I00) to ME, from the Spanish Ministry of Economy and Competitivity (CGL2017-88500-P; AEI/FEDER, EU) to GNF and from the EU-SYNTHESYS programme (NL-TAF- 6894) to LP.

ACKNOWLEDGMENTS

We thank all reviewers and Frontiers editors that helped us edit the articles included in this Research Topic.

REFERENCES

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., et al. (2018). 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

Koenen, E. J., 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

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

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science* 226, 792–801. doi: 10.1126/science.15739260

Soltis, P. S., and Soltis, D. E. (2012). *Polyploidy and Genome Evolution.* Berlin: Springer.

Twyford, A. D., and Ennos, R. A. (2012). Next-generation hybridization and introgression. *Heredity* 108, 179–189. doi: 10.1038/hdy.2011.68

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

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: 11 April 2019
doi: 10.3389/fpls.2019.00345

Cercis: A Non-polyploid Genomic Relic Within the Generally Polyploid Legume Family

Jacob S. Stai[1†], Akshay Yadav[2†], Carole Sinou[3], Anne Bruneau[3], Jeff J. Doyle[4], David Fernández-Baca[5] and Steven B. Cannon[6]*

[1] Interdepartmental Genetics and Genomics Graduate Program, Iowa State University, Ames, IA, United States, [2] Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, IA, United States, [3] Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada, [4] School of Integrative Plant Science, Plant Breeding and Genetics and Plant Biology Sections, Cornell University, Ithaca, NY, United States, [5] Department of Computer Science, Iowa State University, Ames, IA, United States, [6] Corn Insects and Crop Genetics Research Unit, US Department of Agriculture–Agricultural Research Service, Ames, IA, United States

OPEN ACCESS

Edited by:
Marcial Escudero, Universidad de Sevilla, Spain

Reviewed by:
Martin A. Lysak, Masaryk University, Czechia
Jean-Francois Arrighi, Institut de Recherche pour le Développement (IRD), France
Andre Chanderbali, University of Florida, United States

***Correspondence:**
Steven B. Cannon
steven.cannon@ars.usda.gov

† 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: *21 December 2018*
Accepted: *05 March 2019*
Published: *11 April 2019*

Citation:
Stai JS, Yadav A, Sinou C, Bruneau A, Doyle JJ, Fernández-Baca D and Cannon SB (2019) Cercis: A Non-polyploid Genomic Relic Within the Generally Polyploid Legume Family. Front. Plant Sci. 10:345. doi: 10.3389/fpls.2019.00345

Based on evolutionary, phylogenomic, and synteny analyses of genome sequences for more than a dozen diverse legume species as well as analysis of chromosome counts across the legume family, we conclude that the genus *Cercis* provides a plausible model for an early evolutionary form of the legume genome. The small *Cercis* genus is in the earliest-diverging clade in the earliest-diverging legume subfamily (Cercidoideae). The *Cercis* genome is physically small, and has accumulated mutations at an unusually slow rate compared to other legumes. Chromosome counts across 477 legume genera, combined with phylogenetic reconstructions and histories of whole-genome duplications, suggest that the legume progenitor had 7 chromosomes – as does *Cercis*. We propose a model in which a legume progenitor, with 7 chromosomes, diversified into species that would become the Cercidoideae and the remaining legume subfamilies; then speciation in the Cercidoideae gave rise to the progenitor of the *Cercis* genus. There is evidence for a genome duplication in the remaining Cercidoideae, which is likely due to allotetraploidy involving hybridization between a *Cercis* progenitor and a second diploid species that existed at the time of the polyploidy event. Outside the Cercidoideae, a set of probably independent whole-genome duplications gave rise to the five other legume subfamilies, at least four of which have predominant counts of 12–14 chromosomes among their early-diverging taxa. An earlier study concluded that independent duplications occurred in the Caesalpinioideae, Detarioideae, and Papilionoideae. We conclude that Cercis may be unique among legumes in lacking evidence of polyploidy, a process that has shaped the genomes of all other legumes thus far investigated.

Keywords: Cercis, polyploidy, legume family, chromosome evolution, whole-genome duplication, ancestral genome

INTRODUCTION

The legume family, Leguminosae, with approximately 20,000 species, is the third most diverse plant family, after Orchidaceae and Asteraceae (Legume Phylogeny Working Group et al., 2017). The family underwent a rapid radiation shortly after its origin ∼59–64 million years ago (Mya) (Lavin et al., 2005; Bruneau et al., 2008), giving rise to six lineages that have recently been recognized as subfamilies by the international legume systematics community (Legume Phylogeny Working Group et al., 2017). Among those subfamilies, four of them (Papilionoideae, Caesalpinioideae, Detarioideae, Cercidoideae) contain the vast majority of genera and species, while Dialioideae contains 17 genera and 84 species, and Duparquetioideae contains a single genus and species. The four larger subfamilies have been shown (Cannon et al., 2015) to each have been affected by early whole-genome duplications (WGDs): at the base of the Papilionoideae and near the origins of the Cercidoideae, Detarioideae, and Caesalpinioideae – though the precise timing of the WGD(s) in the latter three lineages remains uncertain due to low sampling.

In particular, the WGD status and timing within the Cercidoideae has been uncertain: did a WGD predate the earliest divergences in the family, or did it occur later? Cannon et al. (2015) reported a WGD signal for *Bauhinia tomentosa*, based on comparisons of divergence times of duplicated genes and orthologs based on synonymous substitution distributions (K_s peaks for duplication and speciation) from transcriptome sequence – but no WGD peak was evident for *Cercis canadensis*. This result was inconclusive, however: lack of a WGD peak could have been due to sequence loss or non-recovery for that genus. The genus *Cercis* is sister to the remainder of the Cercidoideae genera (Lewis et al., 2005; Sinou et al., 2009; Wang et al., 2018); we therefore address the question of whether *Cercis* was affected by an early WGD or whether the WGD occurred later in the evolution of the subfamily.

The legumes fall within the Fabidae (rosid 1) clade (Angiosperm Phylogeny Group et al., 2016), and thus were affected by the gamma triplication event that occurred around the time of the origin of the core eudicots, approximately 120 Mya (Jiao et al., 2012). Species such as *Phaseolus* (bean; papilionoid) or *Desmanthus* (bundleflower; caesalpinioid) show evidence of old but independent duplications within the legume family (Cannon et al., 2015). Finding one or more early-diverging legume species without WGD would be of interest because such species could provide important clues to both the structure of the ancestral legume genome and the evolution of species and genomes across this large family.

In the present study, we investigate a new set of genome sequences from the Cercidoideae, Caesalpinioideae, and Papilionoideae, as well as extensive chromosome count data from across the legumes. We also describe results from targeted sequencing of selected genes within the Cercidoideae, to clarify the timing and nature of WGDs affecting the legumes. We present evidence supporting lack of a WGD in the genus *Cercis*, and hypothesize an allotetraploidy event affecting the remainder of the Cercidoideae subfamily.

MATERIALS AND METHODS

Gene Family Construction, K_s Analysis, and Phylogeny Calculation

Gene families include proteomes (complete sets of translated coding sequences – one representative transcript per gene) from fifteen legume species, and five non-legume species – which were used for phylogenetic rooting and evolutionary context. Species and sources are indicated in **Table 1**. We used a custom gene family construction method in order to best capture some challenging features of the phylogeny. Gene family features to account for include early WGDs affecting species in the family – but we wished to avoid an older genome triplication, occurring early in angiosperm evolution. Therefore, we used a combination of homology filtering based on per-species synonymous site changes, comparison with outgroup species, Markov clustering, and progressive refinements of family hidden Markov models (HMMs). The gene families are available at https://legumeinfo.org/data/public/Gene_families/legume.genefam.fam1.M65K/ and associated methods and scripts are available at https://github.com/LegumeFederation/legfed_gene_families although the resources at those locations are focused on papilionoid species rather than on the non-papilionoid species examined in this paper. The same gene families above were used in the analysis in this paper, but with several papilionoid species removed and five other species added (via HMM-search and HMM alignment of the other species to the gene-family HMMs), as shown in **Table 1**. Resources for these gene families are available in Supplementary Materials: **Supplementary Data Sheet S1** (full alignments), **Supplementary Data Sheet S2** (trimmed alignments), **Supplementary Data Sheet S3** (maximum likelihood trees), and **Supplementary Data Sheet S4** (maximum likelihood trees, with same-species terminal pairs reduced to a single representative).

Gene families were generated as follows. All-by-all comparisons of protein sequences for all species were calculated using BLAST (Camacho et al., 2009). Matches were filtered to the top two matches per query, with at least 50% query coverage and 60% identity. For the resulting gene pairs, in-frame nucleotide alignments of coding sequences were calculated, which were used, in turn, to calculate synonymous (K_s) counts per gene pair, using the PAML package (Yang, 2007), with the Nei and Gojobori (1986) method for estimating the numbers of synonymous nucleotide substitutions. The calculation process was driven using the synonymous_calc.py wrapper script (Tang and Chapman, 2018), which additionally uses the packages biopython (Cock et al., 2009), ClustalW2 (Larkin et al., 2007), and PAL2NAL (Suyama et al., 2006). For each species pair, histograms of K_s frequencies were used as the basis for choosing per-species K_s cutoffs for that species pair in the legumes. For most species pairs, the selected peak corresponded with the papilionoid duplication (K_s average of

TABLE 1 | Genome and annotation sources and versions.

Species	Genotype	Assembly	Annot.	Citation	Source
Arachis duranensis	V14167	1	1	Bertioli et al., 2016	PeanutBase
Arachis ipaensis	K30076	1	1	Bertioli et al., 2016	PeanutBase
Cajanus cajan	ICPL87119	1	1	Varshney et al., 2012	LegumeInfo
Glycine max	Williams 82	2	1	Schmutz et al., 2010	Phytozome
Phaseolus vulgaris	G19833	2	1	Schmutz et al., 2014	Phytozome
Vigna radiata	VC1973A	6	1	Kang et al., 2014	LegumeInfo
Lotus japonicus	MG20	3	1	Sato et al., 2008	Phytozome
Medicago truncatula	A17_HM341	4	2	Tang et al., 2014	Phytozome
Cicer arietinum	Frontier	1	1	Varshney et al., 2013	LegumeInfo
Nissolia schottii		1	1	Griesmann et al., 2018	GigaDB
Mimosa pudica		1	1	Griesmann et al., 2018	GigaDB
Chamaecrista fasciculata		1	1	Griesmann et al., 2018	GigaDB
Bauhinia tomentosa		1	1	Cannon et al., 2015	GigaDB
Cercis canadensis		1	1	Griesmann et al., 2018	GigaDB
Prunus persica	Lovell	2	2.1	International Peach Genome Initiative[IPGI], 2013	Phytozome
Cucumis sativus		1	1	Phytozome 12, 2018	Phytozome
Vitis vinifera	PN40024	12X	12X	Jaillon et al., 2007	Phytozome
Arabidopsis thaliana	Col-0	TAIR10	TAIR10	Berardini et al., 2015	Phytozome
Solanum lycopersicum	LA1589	ITAG2.4	ITAG2.4	The Tomato Genome Consortium, 2012	Phytozome

0.6, varying between 0.45 and 0.8; **Supplementary Table S1**). For comparisons between papilionoid species and the four non-papilionoid legume species (*Mimosa pudica*, *Chamaecrista fasciculata, B. tomentosa,* and *C. canadensis*), the selected peak corresponded to the speciation divergence between the pair of species. To accommodate variation in K_s values, the cutoff for each species pair was generally set at 1.5 times the modal K_s value (K_s peak). The set of gene pairs was filtered to remove all pairs with K_s values greater than the per-species-pair K_s cutoff. The resulting set of filtered pairs was used for Markov clustering, implemented in the mcl program (Enright et al., 2002), with inflation parameter 1.2, and relative score values (transformed from K_s values) indicated with the -abc flag. Sequence alignments were then generated for all gene families using MUSCLE (Edgar, 2004). Hidden Markov models (HMMs) were calculated from the alignments using the hmmer package (Mistry et al., 2013), and sequences in each family were realigned to the family that those sequences were assigned to, in order to determine HMM bitscores and calculate a median alignment score for each family. Families were then evaluated for outliers: sequences scoring less than 40% of the median HMM bitscore for the family were removed. The HMMs were then recalculated for each family (without the low-scoring outliers), and were used as targets for HMM search of all sequences in the proteome sets – including those omitted during the initial K_s filtering. Again, sequences scoring less than 40% of the median HMM bitscore for the family were removed. These HMM alignments were then used for calculating phylogenetic trees, after trimming non-aligning characters (characters outside the HMM match states). Phylogenies were calculated using RAxML (Stamatakis et al., 2008), with model PROTGAMMAAUTO, and rooted using the closest available outgroup species.

Calculation of K_s Values and Modal K_s Peaks

Synonymous-site differences (K_s) were calculated by two methods: first, based on gene-pairs derived from the top two matches of genes between or within species, based on blastp sequence searches; and second, based on gene-pairs derived from genomic synteny comparisons and coding-sequence coordinates, provided to the CoGe SynMap service at https://genomevolution.org/coge/ (Haug-Baltzell et al., 2017). In the former case (calculated on top blastp matches), K_s values were calculated using PAML, driven by synonymous_calc.py, by Haibao Tang, available at https://github.com/tanghaibao/bio-pipeline. From the PAML output, the Nei-Gojobori K_s value was used (Nei and Gojobori, 1986). For both approaches (BLAST-based and synteny gene-pair-based), K_s histograms were calculated after filtering for K_s values between 0 and 2. The K_s values and plots are available in **Supplementary Table S1**.

Inference of Consensus Branch Lengths From K_s Peaks

To infer branch lengths for an idealized gene tree from these K_s peak values (**Figure 1D**), modal K_s peak values were read from K_s histograms, with values representing WGD events for a species compared with itself (e.g., in *Phaseolus* with respect to the papilionoid WGD) or orthologous gene separations between species (e.g., between *Phaseolus* and *Cercis*). The modal K_s values were then used to algebraically calculate branch lengths along a gene tree with known species topology and hypothesized duplication history, for the selected species. In these calculations, each branch segment is a variable to be solved, given the observed distances between each terminal (e.g., 0.55 for the phylogenetic path between *Phaseolus* and *Cercis*). Because the internal branch

lengths are not uniquely determinable from the observed K_s path-lengths, several branch lengths were set at 0.01 (based on very short branch lengths observed in both gene trees and species trees): branches subtending the *Chamaecrista* WGD, the papilionoid/caesalpinioid clade, and the *Cercis–Bauhinia* 2 clade. Then, a PHYLIP-format (Felsenstein, 1980) gene tree was manually generated for the represented species, using branch length values from the algebraic calculations.

Methods for Mining for Tree Topologies

To test the order of phylogenetic events, gene trees were evaluated for 14,709 legume gene family trees that contain *Cercis* and/or *Bauhinia* sequences. Python scripts[1] that use the functions from the ETE Toolkit (Huerta-Cepas et al., 2010, 2016) were used to read and analyze the legume gene family trees using the species overlap method (Huerta-Cepas et al., 2007). The species overlap method labels an internal node in a given rooted tree as D (duplication event) or S (speciation event) based on whether there are common species between both partitions corresponding to the two subsequent children nodes. Species-overlap tests were run for trees in which same-species terminal pairs were collapsed (when both branch lengths were less than 0.01), to control for local private gene duplications (**Supplementary Data Sheet S4**).

RESULTS

K_s Peaks From Self-Comparisons of Coding Sequence

Within- and between-species comparisons of rates of synonymous-site changes per synonymous site were evaluated by Cannon et al. (2015) for 20 diverse legume species – including representatives from each of the four largest legume subfamilies. These showed K_s peaks of around 0.3–0.6 in all species except *Cercis*, where only a much older peak of ∼1.5 was seen. Because that work was based on transcriptome sequence for most species, there was some question whether the absence of the peak in *Cercis* might be due to poor sequence quality or sequence non-recovery (although the transcriptome assembly statistics were generally in the same range as for the other species). Recent availability of genome sequences for *C. canadensis*, *C. fasciculata*, *M. pudica*, and *Nissolia schottii*, from Griesmann et al. (2018), provides an opportunity to test K_s and other results with greater rigor. *Chamaecrista* and *Mimosa* fall within the Caesalpinioideae subfamily, and *Nissolia* is in the Papilionoideae subfamily, within the dalbergioid clade, along with peanut (*Arachis*). For K_s analysis in this study, we focus particularly on *Cercis, Bauhinia* (as representatives of the Cercidoideae), *Chamaecrista* (as a representative from the Caesalpinioideae), and *Phaseolus* (as a representative of the Papilionoideae), to investigate evidence for the presence and timing of possible WGDs in these lineages. We include *Phaseolus* to provide an example of a species with high-quality genome sequence and a well-studied, early WGD.

K_s results from genes predicted in the *C. canadensis* ("cerca") and *C. fasciculata* ("chafa") genome assemblies are shown in **Figure 1**, along with genes from *Phaseolus vulgaris* ("phavu") and from *B. tomentosa* ("bauto"; transcriptome-derived). The K_s values were determined both for top BLAST-based gene-pairs between species and within species (e.g., top pairs within *Cercis*). Underlying data for the histograms is available in **Supplementary Table S1**.

There is a clear K_s peak for *Cercis–Bauhinia* at 0.15 and a peak for *Bauhinia* compared with itself at 0.25 (**Figures 1A,C**). Although there are some duplications near 0 in *Cercis* compared with itself, there is no older *Cercis–Cercis* peak as the prominent peak seen in *Bauhinia–Bauhinia* at 0.25. The duplications near 0 in the *Cercis–Cercis* plot are likely due to local gene duplications (as also seen, for example, in the *Phaseolus–Phaseolus* self-comparison in **Figures 1A** vs **1B**), as this signature of recent duplications is absent in the synteny-derived K_s plots in **Figure 2**.

We find the expected strong WGD peak within *Phaseolus* and also for *Phaseolus–Cercis* (at 0.6 and 0.55), respectively, but again, no older peak within *Cercis* compared with itself (**Figure 1B**). The fact that the *Phaseolus–Phaseolus* modal K_s peak is greater than the *Phaseolus–Cercis* peak suggests a much greater rate of mutation accumulation in *Phaseolus* and its progenitors in Papilionoideae than in *Cercis* and its progenitors in Cercidoideae (Cui et al., 2006; Schmutz et al., 2014).

In **Figure 1C**, there is a speciation peak for *Phaseolus–Bauhinia* that is similar to *Phaseolus–Cercis* with the exception that the *Phaseolus–Bauhinia* peak appears slightly "older" than for *Phaseolus–Cercis* (0.6 vs. 0.55), suggesting more rapid rate of mutation accumulation in *Bauhinia* than in *Cercis*.

Figure 1D shows an inferred consensus gene tree, with branch lengths calculated (with approximation) from K_s plots in **Figures 1**, **2** (as described in Methods).

In **Figures 2A–C**, K_s values are derived from gene-pairs within synteny blocks derived from genome comparisons. A major effect of this strategy is to exclude local gene duplications – and to reduce other paralogous matches that can show up as recent duplications – for example, in matches among many members of a recently expanded gene family. This reduction in recent- and locally derived paralogs is evident in K_s counts near zero for "young" (small) K_s values. The sloping K_s histogram seen in **Figure 1** for *Cercis–Cercis* is entirely absent in **Figure 2**. The modal K_s "peak" for *Cercis*, if there is any, is in the range of 1.5–2 – contrasting with the *Cercis–Phaseolus, Cercis–Chamaecrista,* and *Chamaecrista–Phaseolus* peaks of 0.6, 0.5, and 0.7, respectively – indicating that any *Cercis* WGD peak in this data would well predate the legume origin.

Also noteworthy in **Figure 2** is the low modal K_s peak for *Chamaecrista–Chamaecrista* (amplitude of 101, compared with 581 for *Phaseolus–Phaseolus*). This difference in numbers of paralogous duplicated genes could be due to higher rates of gene loss from *Chamaecrista* following WGD early in the Caesalpinioideae. The strong K_s peaks in the orthologous *Chamaecrista – Cercis* comparison and the *Phaseolus – Cercis* comparison suggest that there is nothing systematically wrong with the *Chamaecrista* gene models. Rather, it appears that *Chamaecrista* is more fully "diploidized," with a higher proportion of duplicated genes having reduced to single copies, providing a sufficient basis for discovering correspondences

[1] https://github.com/akshayayadav/clade-based-family-analysis

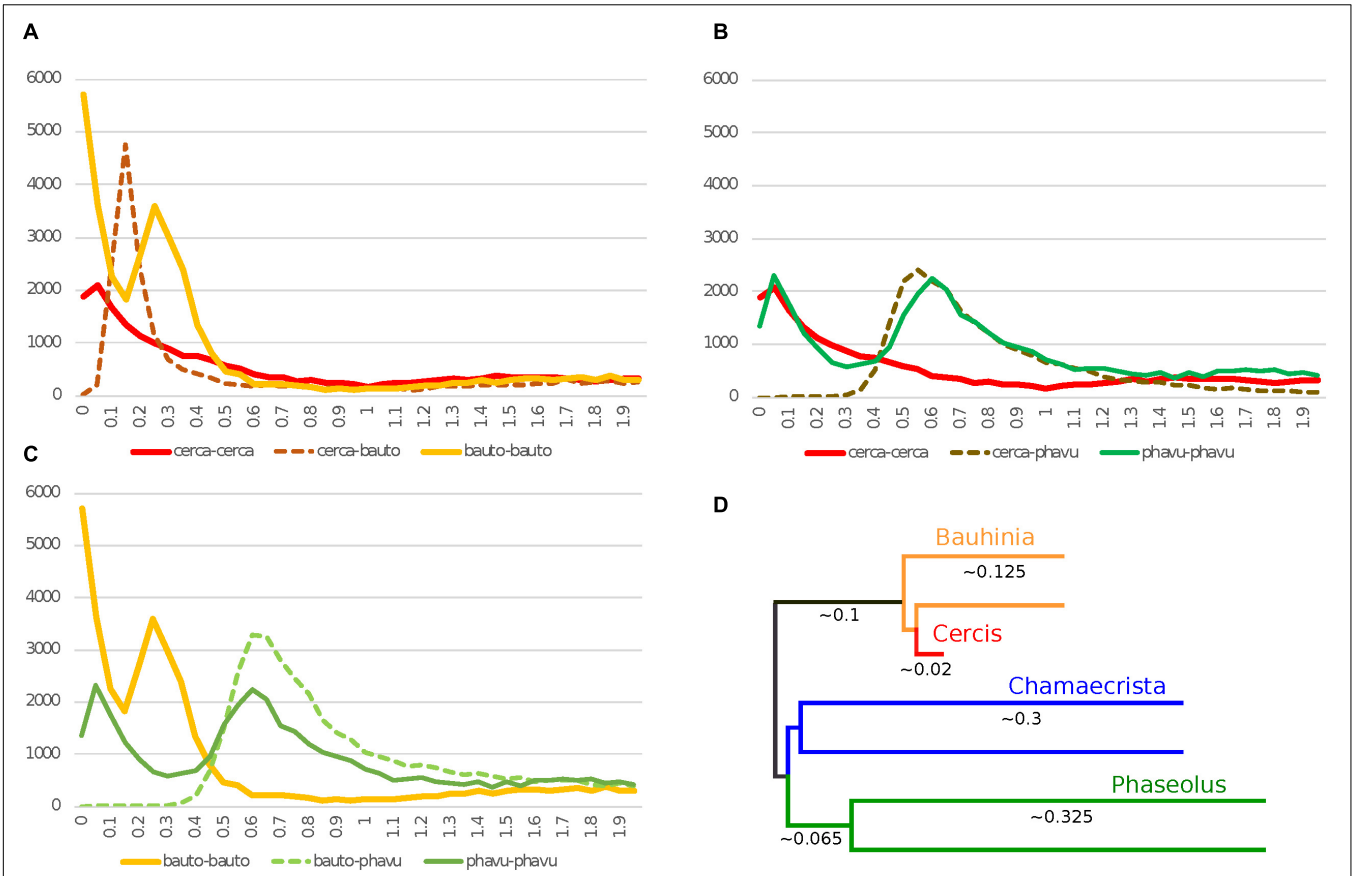

FIGURE 1 | Histograms of K_S values for top gene-pair comparisons for *Cercis canadensis* ("cerca"), *Bauhinia tomentosa* ("bauto"), and *Phaseolus vulgaris* ("phavu"). In K_S plots **(A–C)**, solid lines are for self-comparisons (e.g., for *Cercis* gene-pairs), and dotted lines are for between-species comparisons (e.g., between *Cercis* and *Bauhinia*). The schematic tree in panel **D** is an idealized distance tree in which each OTU represents an "average" gene: either a single copy in *Cercis*, or each of two homoeologs created by unique WGD events in the remaining taxa. Branch lengths are calculated from pairwise modal K_S values in panels **A–C**.

with other species, but erasing much of the WGD signature in a *Chamaecrista* self-comparison. Similar diploidization and interspersed gene losses have been reported in *Medicago truncatula* (Young et al., 2011).

Genomic Synteny Analysis

Given the draft genomic sequence assembly for *Cercis*, it is possible to make synteny comparisons with other legume genome assemblies, as well as assemblies of near outgroups to the legumes. In a synteny comparison of two genomes, a WGD present in one of the genomes and absent in the other should be apparent in a genomic dotplot through the following pattern: starting from a given genomic region in the non-duplicated genome and tracing through the dotplot, one should find matches to two regions in the genome with the WGD; and starting from a given genomic region in the duplicated genome and tracing through the dotplot in the other axis, one should find matches to a single region in the genome that lacks the WGD. This can be described in terms of "synteny depth:" the depth of the duplicated genome should be twice that of the non-duplicate genome.

Because the *Cercis* assembly is still highly fragmented (N50 of 421 kb), synteny depth is difficult to assess visually, but it can be measured computationally. The quota-alignment package (Tang et al., 2011) identifies synteny blocks between two genomes, attempting to match a specified pair of synteny depths or "quotas." For example, if genome B has a WGD that A lacks, then the quota for B relative to A would be 2:1. If the quota is mis-specified as 1:1, then a poor coverage score will result for the duplicated genome, because many potential blocks in genome B will be missed. We also note that in the quota-alignment package, in a genome self-comparison, the trivial self-match is suppressed, so the expected quota for a genome with a single WGD, compared with itself, would be 1:1 rather than 2:2.

We used the quota-alignment package to test a range of quotas for all comparisons among *Cercis, Phaseolus*, and *Prunus*. We also provide the corresponding plots and textual results in **Supplementary Data Sheets S8**, **S9**. There is no evidence for a duplication in *Prunus* since the angiosperm whole-genome triplication (WGT) (Jiao et al., 2012; The International Peach Genome Initiative et al., 2013), and there is a known WGD in *Phaseolus* at around 50 Mya (Schmutz et al., 2014; Cannon et al., 2015), so these should serve as useful comparisons relative to *Cercis*. For *Prunus–Phaseolus*, a quota of 1:1 gives *Phaseolus* coverage of only 63.8% (**Table 2**) vs. 96% for *Prunus*, indicating

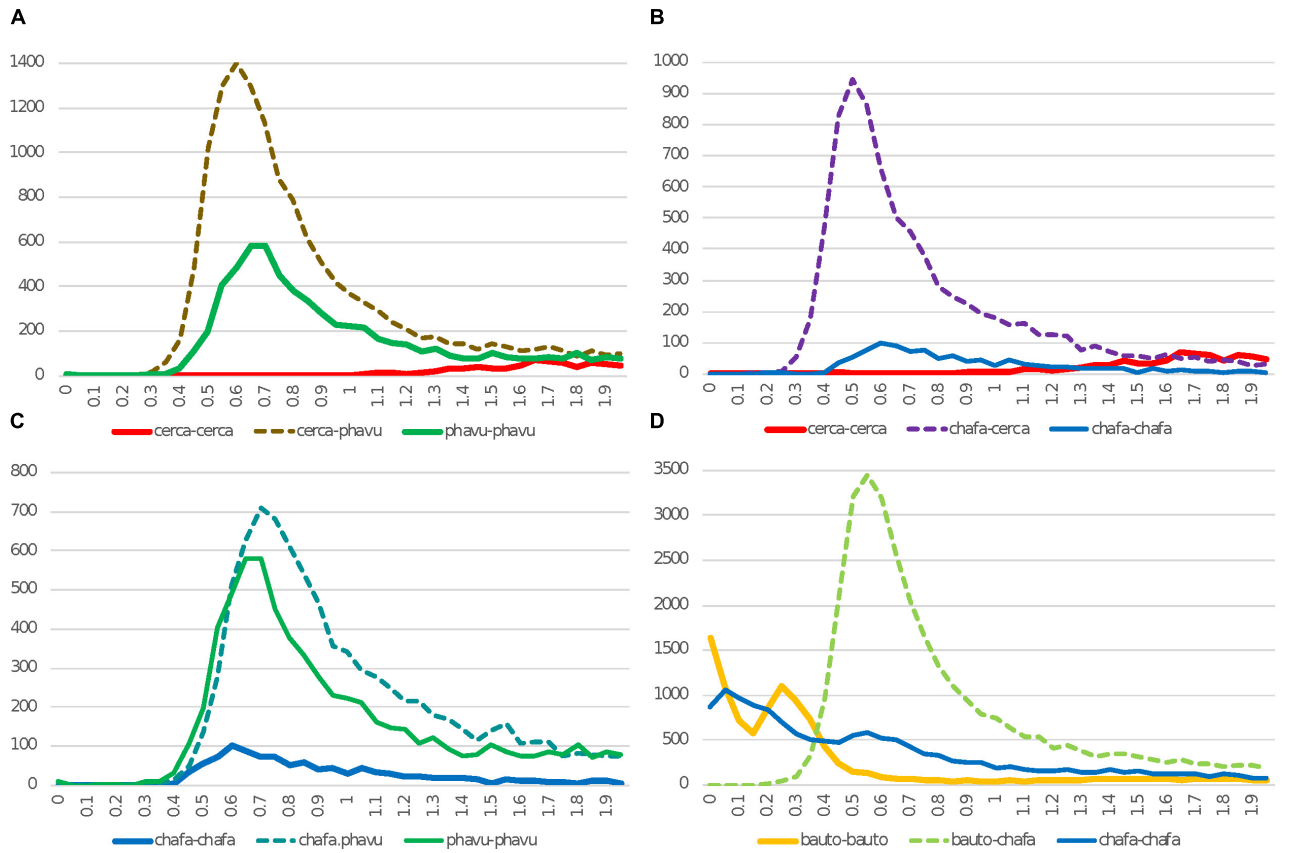

FIGURE 2 | Histograms of K_s values for synteny-based comparisons for *C. canadensis* ("cerca"), *Chamaecrista fasciculata* ("chafa"), *P. vulgaris* ("phavu"), and *B. tomentosa* ("bauto"). In K_s plots, solid lines are for self-comparisons (e.g., for *Cercis* gene-pairs), and dotted lines are for between-species comparisons (e.g., between *Cercis* and *Phaseolus*). This Figure differs from **Figure 1** both in species selection and in method for selecting gene pairs: in **Figure 1**, K_s values are calculated for all top gene pairs, and in panels **A–C**, K_s values are calculated for gene-pairs from synteny features identified from genomic comparisons (panel **D** is an exception: the K_s values are calculated from all top gene pairs, because only transcriptomic sequence is available for *Bauhinia*). The effect of using synteny-based gene pairs for calculating K_s is apparent in the *Chamaecrista* self-comparison plots (chafa–chafa; blue) in panel **B** or **C** (syntenic-based) vs. panel **D** (gene-pair based): in the gene-pair based figures in D, the WGD peak is still evident at ~0.55–0.6, but the signal from more recent gene pairs are also apparent – presumably, as a result of independent, local gene duplications within *Chamaecrista*.

that less than two-thirds of the *Phaseolus* genome has synteny coverage for the identified gene pairs. A quota of 1:2 for *Prunus–Phaseolus* is much better, at 97.4 and 96.8% coverage, respectively. For *Prunus–Cercis*, a quota of 1:1 gives acceptable coverage of 93.4 and 95.2%, respectively; a quota of 1:2 improves the coverage by only about 2% (**Table 2**). For *Phaseolus–Cercis*, the best quota is 2:1, with coverages of 93.3 and 94.7%, respectively. For the self-comparisons for each species, there is notable improvement going from 1:1 to 2:2 (**Table 2**). This is likely due to the ancient angiosperm triploidization (Jiao et al., 2012), which generated three genome copies; the expected number of synteny blocks from any region would then be two (ignoring the trivial self-match).

The K_s peak values derived from gene pairs in the synteny analysis (**Table 2**) are consistent with the synteny depth results – with the *Cercis–Cercis* peak being of comparable age to *Prunus–Prunus* (1.74 and 1.4, respectively), and likely both dating to the angiosperm WGT. In contrast, the peak for *Phaseolus–Phaseolus* is 0.7, consistent with the papilionoid WGD.

Taken together, the synteny and K_s results from **Table 2** indicate that *Cercis* has the same overall WGD depth as *Prunus* and half that of *Phaseolus*, in comparisons among these genomes. In other words, the synteny and K_s evidence supports lack of a WGD in *Cercis*.

Phylogenomic Analyses

To determine duplication events in a phylogenetic context, we constructed gene trees for all legume genes, for fifteen diverse legume species: *Glycine max, P. vulgaris, Vigna unguiculata, Lupinus angularis, Arachis ipaensis, N. schottii, Cicer arietinum, M. truncatula, Lotus japonicus, C. fasciculata, M. pudica, B. tomentosa,* and *C. canadensis*. The first nine of these are from the Papilionoideae (representing the millettioid, genistoid, dalbergioid, and IRLC clades). We also included five non-legume outgroups – using one sequence from each, for each family, in order to provide a rooting for the legume sequences: *Arabidopsis thaliana, Prunus persica, Cucumis sativus, Solanum lycopersicum,* and *Vitis*

TABLE 2 | Synteny coverage for comparisons between the genomes of *Cercis canadensis*, *Phaseolus vulgaris*, and *Prunus persica*, at selected synteny "quotas" (expected coverage depths).

Quotas	*X*	*Y*	K_s peak	Comments
	Cercis	**Cercis**		
q1-1	87.1	87.8	1.74	OK
q2-2	99.9	99.9		BEST
	Phaseolus	**Cercis**		
q1-1	**61.9**	**94.1**	**0.62**	At q1:1, Phaseolus coverage is too low
q2-1	93.3	94.7		BEST
	Prunus	**Cercis**		
q1-1	93.4	95.2	0.92	OK
q1-2	94.1	97.8		little improvement over q1:1
q2-2	99.2	98.6		BEST
	Phaseolus	**Phaseolus**		
q1-1	91.7	92.0	0.70	OK
q2-2	98.9	98.9		BEST
	Prunus	**Phaseolus**		
q1-1	**96.0**	**63.8**	**1.16**	At q1:1, Phaseolus coverage is too low
q1-2	97.4	96.8		BEST
	Prunus	**Prunus**		
q1-1	84.7	84.2	1.40	OK
q2-2	99.6	99.2		BEST

For the comparison between Prunus and Phaseolus (with known WGD histories), the best quota choice is 1:2, corresponding with two synteny blocks in Phaseolus for one in Prunus. Similarly, for the comparison between Cercis and Phaseolus, the best quota choice is 1:2, corresponding with two synteny blocks in Phaseolus for one in Cercis; and for the comparison between Cercis and Prunus, the best quota choice is 1:1, suggesting that neither genome has a recent WGD in its history. The K_s peak values are consistent with this conclusion – with the Cercis–Cercis being of comparable age to Prunus–Prunus (and likely dating to the angiosperm whole-genome triplication). Values in bold highlight cases where quota choices are particularly ill-fitting, and therefore informative as inappropriate models of WGD histories.

vinifera. For convenience, analyses and figures that use sequences from these species use the following abbreviation form to indicate genus and species: the first three letters of the genus and the first two letters of the species epithet, e.g., "glyma" for *G. max*. Gene families were calculated to span the depth of the legume most-recent common ancestor – i.e., avoiding fragmented gene families that split sequences that have a common proto-legume ancestor, and avoiding over-clustered families that include legume sequences that diverged prior to the legume origin. Our method produced 18,543 such families, but for the present analysis, we analyzed the 14,709 families that contain one or more sequences from *Cercis* and/or *Bauhinia*. The set of 14,709 were used for subsequent phylogenomic analyses (**Supplementary Data Sheets S1**–**S4**).

Informal Observations About Patterns in Trees

Gene family trees containing *Cercis* and *Bauhinia* sequences were used to investigate the occurrence of WGD in the most recent common ancestor (MRCA) of the *Cercis* and *Bauhinia* lineages. Although the phylogenomic analysis was likely complicated by uncertainties in phylogenetic reconstructions and by sequence losses or non-recovery, there are clear patterns in the results. We repeatedly see topologies congruent with those in two gene families shown in **Figure 3** (families 31DXWY and 2SH9KY; names from this set of legume gene families were assigned random "license plate" names of six alphanumeric characters). These gene families each show two *Bauhinia* sequences and one *Cercis* sequence in one clade. Both gene families show duplicated sequences for *Mimosa* and *Chamaecrista* (Caesalpinioideae; although in 3A, these do not resolve to a single clade, which may indicate that the duplication occurred very early in the Caesalpinioideae) in the Papilionoideae, there are paired sequences from most species, highlighting the pre-papilionoid WGD (Cannon et al., 2015). In the Cercidoideae clade, there is a curious feature: the duplication that affects *Bauhinia* predates the *Bauhinia*–*Cercis* speciation, and produces the expected two homoeologs in *Bauhinia*, but there is only a single *Cercis* sequence. The full collection of gene trees is available in **Supplementary Data Sheet S3**.

Summaries of Sequence Counts for All Gene Families (Legume Phylogeny Working Group et al., 2017)

To investigate WGDs in the legumes, we analyzed gene counts across all legume gene families. A summary overview of the phylogenomic analysis is shown in **Table 3**, which gives counts of gene families (and trees) having the indicated sequence count for each species (Only selected species are shown in **Table 3**; the counts for all species and all families are given in **Supplementary Table S2**). These are given for two variants of the trees: first (A) for the full, unmodified trees, and second (B) for trees in which similar ($K_s < 0.2$) terminal sequence pairs for a species have been reduced to a single representative, in order to reduce the effect of private, genus-specific WGDs. For example, in **Table 3A**, the first column (glyma / *G. max*) shows the largest number of trees (6531) having two sequences, and the second largest number of trees (3995) having four or more sequences. A count of four for *G. max* would be expected in a gene family in which no gene loss occurred following the two WGDs in the *Glycine* lineage within the period of legume evolution (Schmutz et al., 2010). In **Table 3B**, in which terminal same-species pairs have been reduced to a single representative, the largest number of trees (7951) has one sequence, and the second largest number of trees (4217) has two sequences.

We propose that an indicator of potential older WGDs for a species is obtained by dividing the number of gene family counts for which a species is represented at least twice in the family by the number of family counts for which a species is represented only once. These ratios are given at the bottom of **Tables 3A,B**. For species with a WGD within the period of legume evolution,

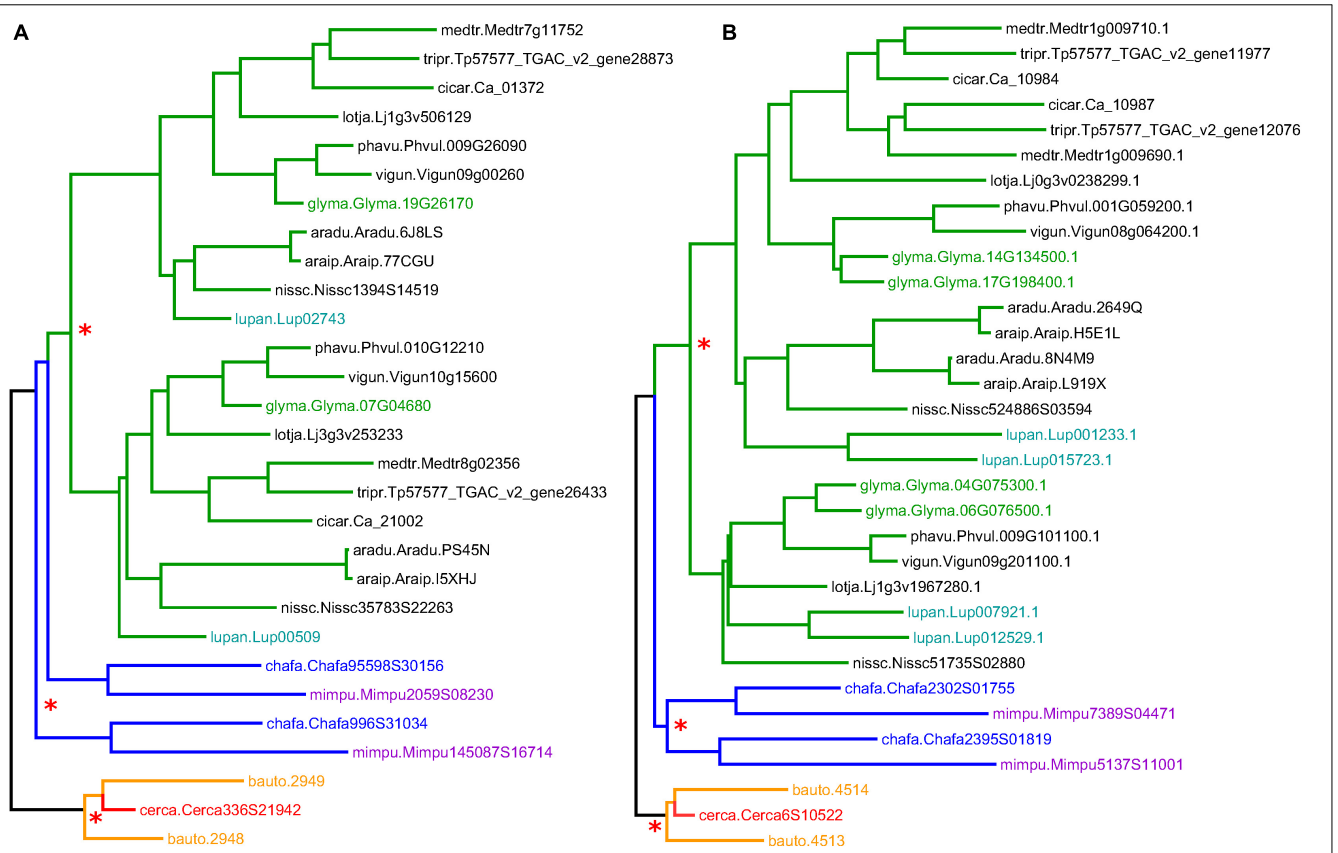

FIGURE 3 | Sample gene trees (for gene families 31DXWY and 2SH9KY; **A** and **B**, respectively), showing clades corresponding to the Cercidoideae (orange and red), Caesalpinioideae (blue and violet), and Papilionoideae (green). Species abbreviations are composed of the first three letters from the genus and the first two letters of the species. Full name correspondences are indicated in the text. Non-legume outgroup sequences are in gray. Red asterisks mark common ancestors of homoeologous sequence pairs. Additional, more recent WGDs within the Papilionoideae are highlighted with colors of the sequence IDs: green for *Glycine max* and turquoise for *Lupinus angustifolius*.

TABLE 3 | Counts of gene families with the indicated numbers of genes per family.

count	glyma	phavu	aradu	Nissc	medtr	tripr	lotja	chafa	mimpu	bauto	cerca
(A) Counts for original full trees.											
0	553	826	2264	1425	1001	1252	1873	2558	3859	4066	1557
1	1933	8748	7761	8472	8141	8255	7602	7894	6432	5921	10567
2	6531	3981	3390	3656	3545	3429	3444	3178	2858	2570	1708
3	1697	716	752	681	984	957	1138	591	846	1130	437
≥ 4	3995	438	542	475	1038	816	652	488	714	1022	440
≥2/ = 1	632%	59%	60%	57%	68%	63%	69%	54%	69%	80%	24%
(B) Counts for trees with terminal recent pairs per species are reduced to a single representative.											
0	553	826	2265	1427	1003	1254	1873	2559	3860	4067	1558
1	7951	9034	7907	8815	8806	8878	9018	8353	7934	7475	10988
2	4217	3911	3396	3621	3443	3285	3066	2970	2160	2362	1564
3	1163	616	707	545	798	791	534	484	430	546	342
≥ 4	825	322	434	301	659	501	218	343	325	259	257
≥2/ = 1	78%	54%	57%	51%	56%	52%	42%	45%	37%	42%	20%

Numbers for a given count (left-hand column) are the numbers of families with counts per species for the given count categories. This Table gives counts for the full gene families, and counts for gene families in which similar (K_s < 0.2) terminal sequence pairs for a species have been reduced to a single representative, in order to reduce the effect of private, genus-specific WGDs. Ratios of counts are given below each table: for a given species, the number of families with one sequence in the family over the number of families with two or more sequences in the family. This ratio provides an indication of possible whole-genome duplications present for that species.

a relatively larger number of families should have two or more sequences. The most dramatic ratio is for *Glycine* (632%; i.e., 6.3 × the naïve expectation) – which has two WGDs in its legume history (pre-papilionoid and a much more recent *Glycine*-specific duplication). For the unreduced trees (1A), all other species have ratios greater than 50% except for *Cercis*, with 24%. For the reduced trees (with collapsed terminal same-species clades), the ratios are somewhat lower for all species: 42–78% for all species except *Cercis*, with 20%. We interpret these results as evidence for WGD in all of the represented legume species except *Cercis*.

TABLE 4 | The types of monophyletic groups containing different numbers of *Cercis* and *Bauhinia* sequences.

# of *Cercis* seqs. in clade	# of *Bauhinia* seqs. in clade	total # of clades detected	# of clades labeled as duplication at MRCA	percent of duplication clades
≥ 2	≥ 2	249	212	85%
≥2	1	425	183	43%
1	≥ 2	3205	2036	63%

For example, there are 425 clades with ≥ 2 Cercis sequences and 1 Bauhinia sequence. The last column indicates the proportion of clades with a duplication pattern consistent with WGD having occurred prior to the Cercis–Bauhinia speciation, e.g., [B,(B,C) or (C,(B,C)], as opposed to a speciation pattern, e.g., [(B,B),C] or [B,(C,C)].

Mining for Tree Topologies Within the Cercidoideae

To infer the relative timing of gene duplications relative to speciations, we mined legume gene phylogenies for topological patterns expected to be produced by these events. Monophyletic groups were detected from a set of 14,709 families containing at least one sequence each from *Cercis* and *Bauhinia* (**Figure 4** and **Table 4**). The MRCA node for each clade containing *Cercis* and *Bauhinia* was labeled either as D (for a duplication event) or S (for a speciation event), based on whether there are common species between both partitions corresponding to the two subsequent children nodes. For example, considering clades with two sequences from each of *Bauhinia* and *Cercis*, [(B,C),(B,C)] would be labeled D while [(B,B),(C,C)] would be labeled S (**Figure 4**) The species overlap method has been previously used to study evolutionary relationships of human proteins with their respective homologs in other eukaryotes (Huerta-Cepas et al., 2007). We considered three types of monophyletic groups varying by number of *Cercis* and *Bauhinia* sequences: clades containing ≥ 2 *Cercis* and ≥ 2 *Bauhinia* sequences, clades containing exactly 1 *Cercis* and ≥ 2 *Bauhinia* sequences, and finally clades containing exactly 1 *Bauhinia* and ≥ 2 *Cercis* sequences. The proportions of clades out of the total number of clades, for all the three types, that were labeled as D at the MRCA node were also calculated. Species-overlap tests were run on trees in which very recently derived same-species terminal pairs were collapsed (when both branch lengths were less than 0.01), to control for local private gene duplications.

There are approximately tenfold more trees with one *Cercis* and two or more *Bauhinia* sequences than with one *Bauhinia* and two or more *Cercis* sequences (**Table 4**; 425/3205 and 183/2036). We interpret this result (preponderance of the 1 *Cercis*, ≥ 2 Bauhinia pattern) as evidence for WGD in *Bauhinia* but not *Cercis*. Further, of the clades with two or more *Bauhinia* sequences and one *Cercis* sequence, most (63%) of these have *Cercis* nested within the clade: 2036 of the total clade count look like [(B,C),B] rather than [(B,B),C] – the former likely resulting from a duplication of *Bauhinia* prior to speciation, and the latter resulting from speciation followed by duplication of *Bauhinia*. This result might seem nonsensical (duplication predating the *Cercis–Bauhinia* speciation, yet not affecting *Cercis*), but it would be consistent with allopolyploidy – with a *Cercis* progenitor having contributed one of the subgenomes in the allopolyploidy event that gave rise to *Bauhinia* and all other species in

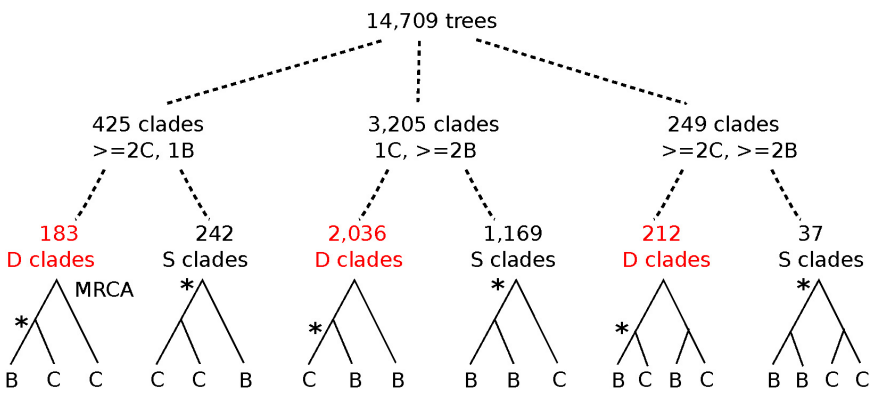

FIGURE 4 | Graphical depiction of tree-mining results for topologies in the Cercidoideae. From 14,709 family trees with *Cercis* and *Bauhinia* sequences, clades with ≥2 *Cercis* and one *Bauhinia* sequence were 7.5 times more common than clades with 1 *Cercis* and ≥2 *Bauhinia* sequences (425 vs. 3205 clades, respectively). Of the latter (more frequent) clade configuration, cases with [**(C,B)**,**B**] are 1.74 times more common than cases with [**(B,B)**,**C**] (2036 vs. 1169 clades, respectively). In the first of these patterns, [**(C,B)**,**B**], the MRCA node of the clade is labeled as a Duplication by the "species overlap" algorithm (see section "Materials and Methods" for description) – meaning that a the MRCA is inferred as due to a gene duplication event rather than a speciation-derived orthology event. Asterisks mark nodes where orthologous genes derive from speciation. Also see **Table 4** for counts and percentages.

the rest of the Cercidoideae clade (elaborated further in the section "Discussion").

Gene Duplication Patterns Across Diverse Species in the Cercidoideae

To determine gene duplication patterns for species in the Cercidoideae, we take advantage of the well-conserved *CYCLOIDEA*-like TCP genes, which have been used both for phylogenetic inference and for studies of evolutionary development in the legumes (Citerne et al., 2003, 2006). Using two sets of degenerate PCR primers that preferentially amplify two classes of *CYCLOIDEA*-like TCP genes in the legumes (Citerne et al., 2003), Sinou and Bruneau (pers. comm.) amplified *CYCLOIDEA*-like genes from 114 species in Cercidoideae. These span all twelve genera in this subfamily. A phylogeny from a subset of these sequences is shown in **Figure 5** – with sequences from each genus included but omitting some species from well-represented genera (see **Supplementary Data Sheet S5** for the phylip-format phylogeny and SD08 for the sequence data and accessions).

A feature readily apparent in the phylogeny is its division into three clades: one with sequences marked "CYC1" (salmon), one with sequences marked "CYC2" (orange), and one unlabeled (red) (**Figure 5**). Most species have two representatives in the phylogeny: one in the CYC1 clade and one in the CYC2 clade – except in *Cercis* (three species), for which only one sequence was amplified (or recovered from the genome assembly, in the case of *C. canadensis*). Although the favored topology places *Cercis* sequences sister to sequences from other Cercidoideae, bootstrap support for this relationship is weak. Alternative resolutions thus are not ruled out, including placement of the *Cercis* clade sister to either CYC1 or CYC2. This would be consistent with the pattern observed in the trees in **Figure 3**, i.e., [(C,B1),B2] – and would be consistent with a model of allopolyploidy (see section "Discussion").

Chromosome Counts Across the Legume Phylogeny

Phylogenetic and chromosome count data can be combined in order to explore chromosomal evolution across the legumes. We combined the extensive *matK*-based phylogeny from the LPWG (Legume Phylogeny Working Group et al., 2017), with count data from the Chromosome Counts Database (CCDB version 1.45) (Rice et al., 2015). The CCDB contains 27,947 count reports for legume species, spanning 477 genera. For many genera, there are numerous reports; for example, *Acacia* has 472 reported counts across 152 species. We determined the modal gametic chromosomal count value, "*n*," for each genus (for example, in *Acacia*, the modal count is $n = 13$, of the 152 species with counts, 71% have $n = 13$). See **Supplementary Table S4** for count details. We then displayed these modal counts on the species phylogeny, using one species as the representative for each genus in the phylogeny.

In **Figures 6**, **7**, a partially collapsed phylogeny has been annotated and summarized for ease of presentation. Count details for each species and genus are given in **Supplementary Table S4**; an image of the full tree with count data is in **Supplementary Data Sheet S6**; and the PHYLIP-format tree file is in **Supplementary Data Sheet S7**. Some particularly well-represented clades have been collapsed; for example, the mimosid clade contains 47 species with chromosomal counts; these have been collapsed in **Figure 7**, and the overall modal count for that clade is presented as an annotation (the mode for the chromosomal count is $n = 14$ for the mimosoid clade within the Caesalpinioideae). See **Table 5** for counts in each clade.

At the subfamily level, the modal chromosome counts are generally unambiguous, with the exception of the Papilionoideae, with a more complex pattern of chromosome counts. The Papilionoideae, being an unusually large subfamily (containing ∼13,800 species in that subfamily and more than 70% of legume species; Cardoso et al., 2012), has been treated in a separate analysis (Ren et al., 2019). However, we note here that the groups sister to the large crown clades of papilionoid species, e.g., *Swartzia, Myroxylon,* and *Cladrastis*, have 13 and 14 as the most frequent counts (**Figure 6** and **Table 5**). The clades of the crown group generally have lower counts: 11 for *Amphimas, Holocalyx, Andira* dispersed along the grade with the genistoid, dalbergioid, and baphioid clades. Among the remaining papilionoid clades (containing the majority of species in the subfamily), chromosome counts are varied, but are generally in the range of 7–11 chromosomes.

The Caesalpinioideae has generally clear count patterns: 14 for the large mimosoid clade and 12–14 for the remaining, early-diverging taxa (**Table 5**). Across 73 genera with counts in the Caesalpinioideae, 66 have modes at $n = 12$, 13, or 14 (14, 35, 17, respectively – combining "early" and "mimosoid" in **Table 5**). There are some intriguing exceptions, however; for example, *Calliandra* and *Chamaecrista* and have $n = 7$–8, despite being nested in clades with $n = 13$ or 14 – apparently indicating chromosomal fusions or reductions of some sort; and other genera such as *Neptunia* and *Leucaena*, have $n = 28$ and 52, respectively, suggesting ploidy increases from $n = 14$ and 13.

For the Dialioideae, five of six genera with count data have $n = 14$. For the Detarioideae, 19 of 23 genera with count data have $n = 12$. For the Cercidoideae, four genera (*Bauhinia, Piliostigma, Griffonia, and Adenolobus*) with count data have $n = 14$, and only *Cercis* has $n = 7$. The nearest outgroup species to the legumes may also be informative. *Quillaja saponaria* (Quillajaceae) which shows evidence of a WGD (via transcriptome K_s data; Cannon et al., 2015), has $n = 14$. Another near outgroup, *Suriana maritima* (Surianaceae), has $n = 9$; its WGD status is not known directly, though it lacks duplication in any of its CYC-like genes (Zhao et al., 2019).

Genome Sizes in the Cercidoideae

Roberts and Werner (2016) report an average of $2C = 0.751$ pg for 30 accessions across 9 *Cercis* species. Using the conversion ratio of 1 pg = 978 Mb (Dolezel et al., 2003), this gives a *Cercis* genome size estimate of $1C = 0.78$ pg * (978 Mb / 1 pg) / 2 = 367 Mbp. This compares with reported 1C genome sizes for several *Bauhinia* species: 573 Mbp for *B. purpurea*; 613 Mbp for *B. tomentosa*, and 620 Mbp for *Lysiphyllum hookeri* (formerly *B. hookeri*) (Bennett and Leitch, 2005). These values are ∼1.5

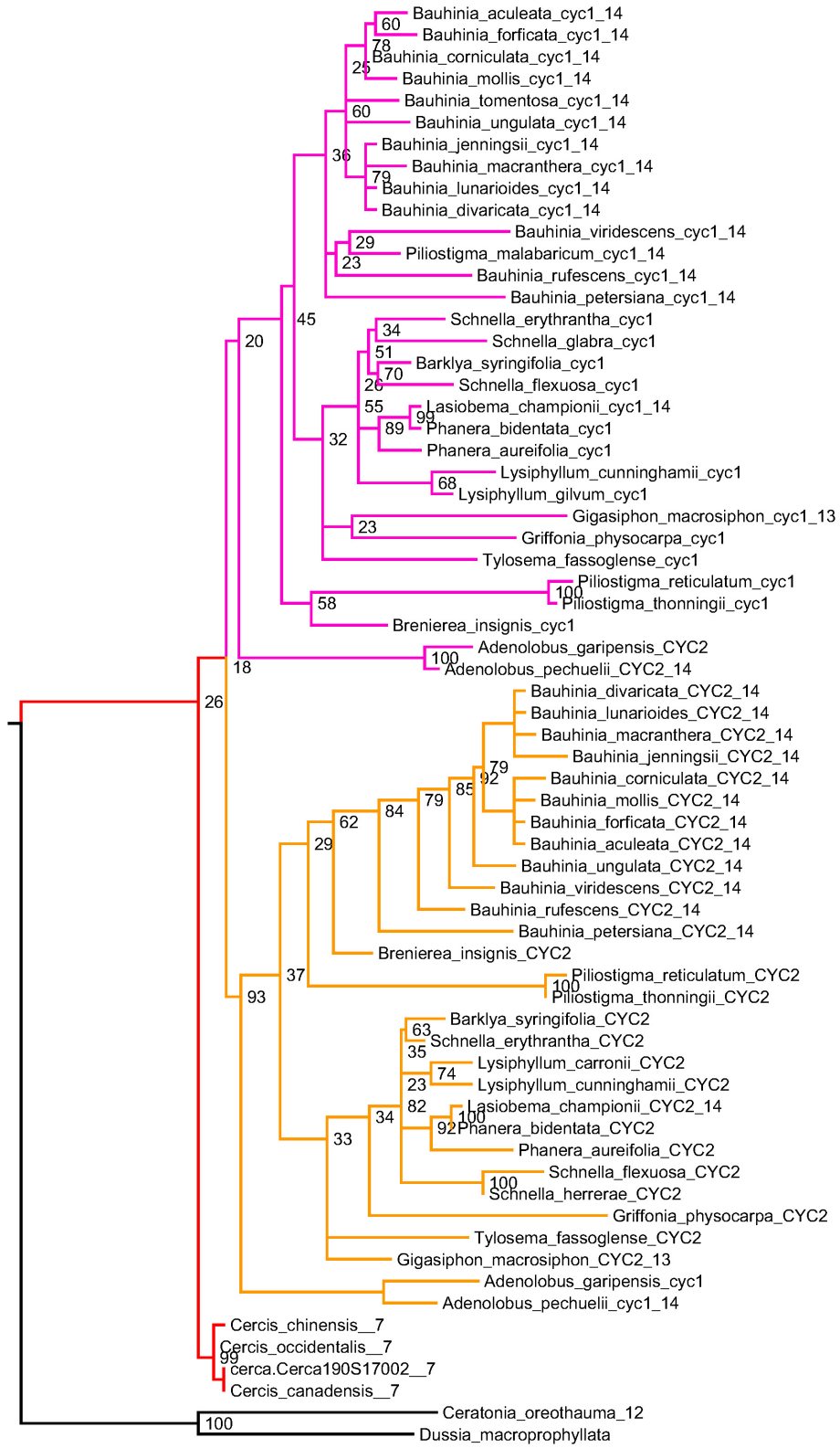

FIGURE 5 | *CYCLOIDEA* gene tree, for species in subfamily Cercidoideae. For all species but *Cercis* (red), there are two gene copies: in the clades labeled "CYC1" (pink) and "CYC2" (orange). Where chromosome counts are available, the haploid count is indicated at the end of the sequence identifier. These values are 7 for the three included *Cercis* species, and 14 for all other species for which counts have been determined within the Cercidoideae, save *Gigasiphon macrosiphon*,

(Continued)

FIGURE 5 | Continued
which has 13. For *C. canadensis*, one sequence has been amplified using PCR and one sequence (Cerca190S17002) comes from the genomic assembly. One of several possible rootings is shown (with bootstrap support values indicated), based on comparison with *CYCLOIDEA* orthologs from *Ceratonia oreothauma* (carob relative, from the Caesalpinioideae) and *Dussia macroprophyllata* (an early-diverging species from the Papilionoideae).

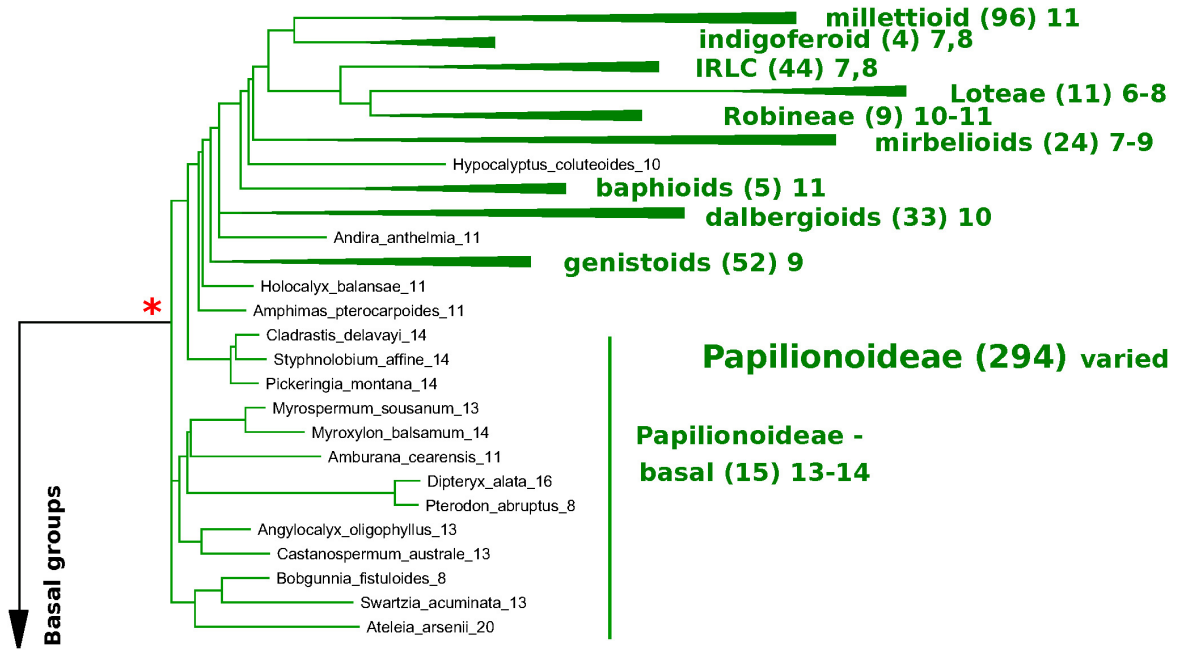

FIGURE 6 | Papilionoid portion of the *matK*-based species phylogeny for representative species in the legumes, with chromosome count data (**Figures 6**, **7**). *matK*-based species phylogeny for representative species in the legumes (derived from Legume Phylogeny Working Group et al., 2017), with chromosome count data. Only species for which chromosome counts are available are shown, with the exception of the Cercidoideae (**Figure 7**), where additional species are shown for context in that subfamily. Chromosomal counts are given as the mode for the indicated genus, where there are differences in the genus. Some particularly well-represented clades have been collapsed and are represented by a colored triangle. The number of genera with counts is given in parentheses – for example, 96 genera are represented in the triangle representing the millettioid clade (top of **Figure 6**), and 47 genera are represented in the triangle representing the Mimosoid clade (top of **Figure 7**). Red asterisks indicate polyploidy events – either known (e.g., Papilionoideae) or hypothesized (e.g., Dialioideae).

to ∼1.6 times larger than *Cercis* – which is consistent with the *Bauhinia* genomes having doubled relative to *Cercis* (followed by moderate increase in *Cercis* and/or decrease in *Bauhinia* – or a situation of an allopolyploid *Bauhinia* being derived from two genomes of different sizes – one contributed by a *Cercis* progenitor and one presumably now extinct). A size of 381 Mbp for *Cercis* is also small relative to other reported legume genomes; for example, the estimated sizes of *L. japonicus, M. truncatula, P. vulgaris,* and *C. arietinum*, respectively, are 472–597 Mbp, 465–562 Mbp, 587–637, 738–929 (Arumuganathan and Earle, 1991; Sato et al., 2008; Bennett and Leitch, 2011; Varshney et al., 2013; Tang et al., 2014). Indeed, in comparison with genome size reports for 722 legume species and 84 genera from the Kew *C*-value database (Bennett and Leitch, 2012), the *Cercis* estimate of $n = 367$ Mbp would be smaller than all but one other legume genome (*Lablab niger* also has an estimated size of 367 Mbp). For all reported legume genera (taking median value per genus where values are available for multiple species in a genus), the average haploid genome size is 1,424 Mbp and the median is 1,157 Mbp (**Supplementary Table S5**).

DISCUSSION

This study examines evidence regarding ploidy in the legume family, particularly focusing on subfamily Cercidoideae. What motivates this focus is the hypothesis that *Cercis,* sister to the remainder of the Cercidoideae, has no history of polyploidy – which may be in contrast to all other legume species. This would make *Cercis* valuable as a genomic model for the legumes, and would also help to clarify histories of chromosome evolution throughout the rest of the large and diverse legume family. Specifically, if *Cercis* did not undergo a WGD relative to the common ancestor of legumes, and if the ancestors of other lineages in the Cercidoideae, Dialioideae, Detarioideae, Caesalpinioideae, and Papilionoideae did, then the legume clade as a whole is not fundamentally polyploid relative to its sister taxa. Combined with evidence that the papilionoid WGD affects all papilionoid species but does not extend to species in the caesalpinioid or detarioid subfamilies (Cannon et al., 2015), the necessary inference is that there must have been multiple, independent events: at a minimum, one in the Cercidoideae

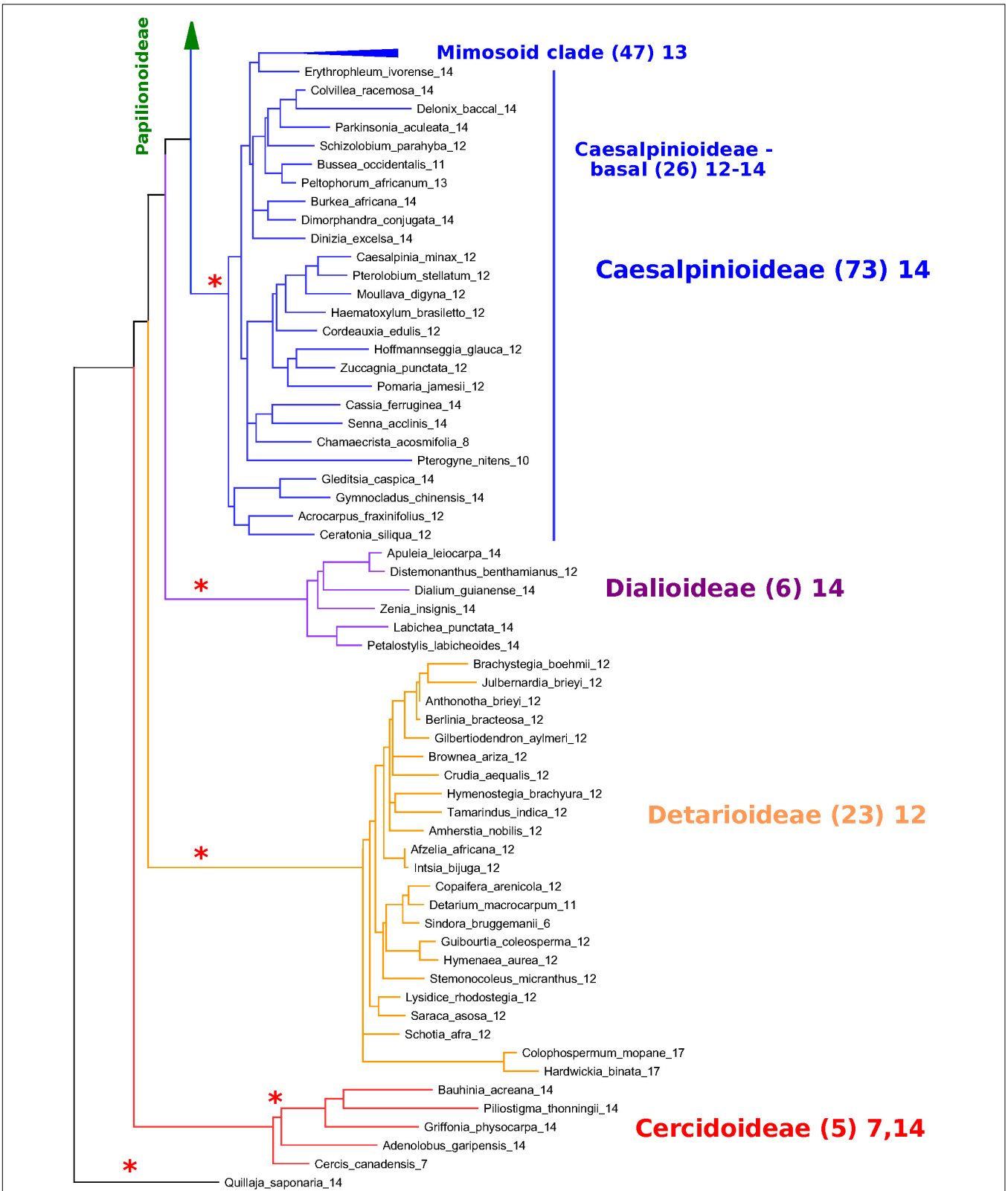

FIGURE 7 | Non-papilionoid portion of the *matK*-based species phylogeny for representative species in the legumes, with chromosome count data. **Figure 7** extends **Figure 6**; see description under **Figure 6**. The relative placements of the subfamilies are uncertain, with the Cercidoideae and Detarioideae, best considered as a polytomy, given current phylogenetic resolutions (Legume Phylogeny Working Group et al., 2017). *Indicate polyploidy event – either known (e.g., Papilionoideae) or hypothesized (e.g., Dialioideae).

TABLE 5 | Counts of genera with indicated haploid (gametic) chromosome numbers, by subfamily or clade.

Clade\Count	6	7	8	9	10	11	**12**	**13**	**14**	15	16	> 16	total	frequent
Papilionoid – derived	4	21	**57**	**36**	**39**	**77**	6	0	5	0	6	27	278	8–11
Papilionoid – grade	0	0	0	0	0	**3**	0	0	0	0	0	0	3	11
Papilionoid – early	0	0	2	0	0	1	0	**4**	**4**	0	1	1	13	13–14
Caesalp – mimosoid	0	0	1	0	0	0	1	**31**	5	0	0	3	41	13
Caesalp – early	0	0	1	0	1	1	**13**	**4**	**12**	0	0	0	32	12–14
Dialidoideae	0	0	0	0	0	0	1	0	**5**	0	0	0	6	14
Detarioideae	1	0	0	0	0	1	**19**	0	0	0	0	2	23	12
Cercidoideae	0	**1**	0	0	0	0	0	0	**4**	0	0	0	5	7,14

Each cell (except for the count summaries in the last three columns) contains the number of genera with a chromosome count indicated (column), for that clade (row). For example, in the Caesalpinioideae (which includes the mimosoid clade), 31 genera have a chromosome count of 13. (For most genera, all species have the same chromosome count, but where count differences are reported in the literature, the modal value is used for the genus). For each clade, the most frequent chromosome count is highlighted in bold, and the most frequent count values are listed on the right. Species and count details are given in ***Supplementary Table S2****.*

and another in the Papilionoideae – and our findings here are also consistent with our previous conclusion of independent polyploidy events early in the Caesalpinioideae and Detarioideae (Cannon et al., 2015). We have no information about ploidy in the monogeneric Duparquetioideae; and it is not known directly whether species in the Dialioideae experienced a WGD, though chromosome counts of 12–14 in Dialioideae are consistent with the hypothesis that they too are polyploid.

The cumulative evidence that *Cercis* lacks a legume-era WGD is substantial. Recapping:

- In K_s plots (**Figures 1**, **2**), there is no peak indicating WGD in *Cercis* – particularly, in plots derived from synteny comparisons. In contrast, such peaks are clearly evident in diverse legume lineages including *Phaseolus, Bauhinia,* and *Chamaecrista*. While there is no such peak in the *Cercis* self-comparison, there are clear peaks in comparisons of *Cercis* to each of the other species examined, indicating that the lack of K_s peak is not due to something essentially wrong with gene-calls in *Cercis* (the gene calls have homologs with the comparison legume species, and those homologs can be aligned in-frame with those homologs, giving reasonable K_s results).
- In genomic synteny comparisons between *Cercis*, *Phaseolus*, and *Prunus* (the latter two with known duplication histories), the duplication status of *Cercis* looks like that of *Prunus* rather than *Phaseolus* – i.e., lacking a WGD in the timeframe of the fabidae.
- In phylogenomic analyses of 14,709 gene-family trees (**Table 3**), sequence counts aggregated across all trees show a pattern consistent with at least one WGD in each species examined except *Cercis*. Examining the proportion of gene families with two or more sequences for a species to families with only one sequence, all species examined have a ratio ranging from 54 to 80% (and 632% for *G. max*, which had an additional recent WGD), in contrast to 24% for *Cercis*. For comparison, this ratio is 69% in the set of 177 conserved collinear genes in the triplicated *B. oleracea* genome segments identified by Town et al. (2006).
- Mining the gene families for phylogenetic topologies within the Cercidoideae (**Table 4**), the overwhelming majority of clades have a pattern of two *Bauhinia* sequences to one *Cercis* sequence (roughly tenfold more frequently than the other options combined).
- Diverse species within the Cercidoideae all show a pattern of duplicated *CYCLOIDEA*-family genes, with the exception of *Cercis*, which has only one *CYCLOIDEA* gene – whether assayed through amplification with degenerate primers for *CYCLOIDEA*, or through gene prediction in the *Cercis* genomic sequence (**Figure 5**). All phylogenetic analyses (whether based on plastid or nuclear sequences) resolve *Cercis* as sister to the remainder to Cercidoideae, in line with a WGD after the split with *Cercis* (although rooting in **Figure 5** is uncertain, so *Cercis* could group with one or the other of the *CYCLOIDEA* gene forms in the gene family).
- A survey of chromosome count data for 477 legume genera, examined in a phylogenetic context (**Figure 7**, **Table 5**, **Supplementary Table S4**, and **Supplementary Data Sheets S6, S7**), shows a pattern consistent with WGDs affecting all subfamilies and most genera – with the exception of *Cercis* itself. Models in which most legumes are polyploid have been proposed in earlier studies (Goldblatt, 1981; Doyle, 2012), on the basis of chromosome numbers. In the Cercidoideae, the most frequent chromosome count is $n = 14$ for most species, but 7 in *Cercis*; in the Detarioideae, the modal chromosome count is 12; in the Dialioideae, the modal count is 14; in the Caesalpinioideae, the modal count is 14; and in the Papilionoideae, the modal count for early-diverging genera (e.g., *Swartzia*, *Angylocalyx*, *Cladrastis*), the most common counts are 13 and 14. Crown-group clades have highly variable counts (generally in the range of 7–11 chromosomes), so we hypothesize a doubling from 7 to 14 leading to the papilionoid origin, then a reduction from 14 to lower numbers for crown-group clades (dalbergioids, baphioids, mirbelioids, Robineae, Loteae, IRLC, indigoferoid, and millettioid).
- Genome sizes in the Cercidoideae are consistent with WGD in *Bauhinia* and not *Cercis*. The *Cercis* genome is approximately 367 Mbp, while values for *Bauhinia* species range from 573 to 620 Mbp. A *Cercis* genome size of 367

Mbp is tied for smallest in the legume family, and is less than a third the median reported genome size of 1,157 Mbp, across 84 legume genera. We note this result with a caveat, however, that genome sizes can be highly variable, even within a single genus – affected by mechanisms such as bursts of transposon expansions – e.g., variations in *Nicotiana* (Leitch et al., 2008) or in *Aeschynomene* (Brottier et al., 2018).

Further analyses of evolutionary changes due to the differing WGD status between *Cercis* and other legumes will be of interest – both at the fine scale (e.g., determining the fate of duplicated genes in various lineages, relative to *Cercis*) and at larger structural scales (e.g., determining structural changes in chromosomes following several independent WGD events) These comparisons would benefit from improved assemblies and annotations, spanning a broader range of legume clades. For example, we expect both *Chamaecrista* (as a nodulator in the Mimosoideae) and *Cercis* (as an early-diverging non-nodulator) to be useful in better understanding the origin and evolution of nodulation symbioses – as investigated in several recent papers (Battenberg et al., 2018; Griesmann et al., 2018; van Velzen et al., 2018).

An initially puzzling result from our analysis was the fact that the K_s peak for the *Bauhinia* self-comparison (*Bauhinia–Bauhinia*) appears significantly "older" than the *Bauhinia–Cercis* speciation peak, at 0.25 and 0.15, respectively (**Figure 1A**). Similarly, most gene tree topologies (63%) that have two or more *Bauhinia* sequences and one *Cercis* sequence (**Table 4**, row 3) have a configuration of (B,(B,C)), indicating duplication prior to speciation – in contrast to what might be expected given a simple model of *Cercis–Bauhinia* speciation followed by WGD in *Bauhinia*. In the latter case, the expected pattern would be [(B,B),C] – which is observed in the minority of cases (37%). We note that an apparent speciation pattern may be due either to a WGD or to local, private duplications. Private duplications are common in plant genomes. For example, in *M. truncatula*, more than a third of paralogs are derived from local duplications (Young et al., 2011). However, local duplications tend to be evident in K_s plots as a recent peak, with maximum near zero – as is seen, for example, in the *Phaseolus–Phaseolus* comparison in **Figure 1**. This is the typical pattern described by Lynch and Conery (2000) for eukaryotes generally. The results of our phylogenetic pattern-mining tests are consistent with what we observe (albeit anecdotally) in visual inspection of many trees, exemplified by **Figure 3**, in which there is a duplication of the *Bauhinia* paralogs in both trees, apparently followed by orthologous split between one of the *Bauhinia* sequences and the *Cercis* sequence.

A model that could accommodate the K_s and tree-topology results is one of allopolyploidy, in which a progenitor of *Cercis* speciated to give another (perhaps now-extinct) diploid species (**Figure 8A**). These species diverged for some time, and then the two species contributed their genomes to a new allopolyploid species that was the progenitor of the remaining Cercidoideae. Following allopolyploidy, the two lineages (diploid *Cercis* and polyploid *Bauhinia*) would then have proceeded to diverge and diversify – *Cercis* more slowly and the remaining species in Cercidoideae more rapidly. The current gene family view would then be as observed in e.g., **Figure 3**, or in the model in **Figure 8B**.

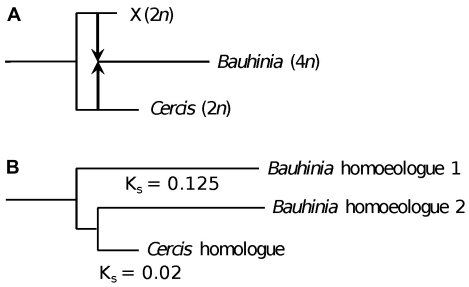

FIGURE 8 | Allopolyploid origin of *Bauhinia*. **(A)** Species history, showing divergence between two diploid (2*n*) species: (1) the ancestor of *Cercis* and (2) a second species that became extinct ("X"). At some point after the species divergence, the two diploid species hybridized (arrows), followed by genome doubling to produce the allopolyploid (4*n*) ancestor of *Bauhinia* (and other Cercidoideae). **(B)** Representative gene tree sampled from *Bauhinia* and *Cercis*, showing the relationships of the single homologous gene in *Cercis* to the two homoeologs in allopolyploid *Bauhinia*. The *Bauhinia* homoeolog 2, contributed by the *Cercis* ancestor, is sister to the *Cercis* gene. The Cercis gene has a K_s of ∼0.145 compared with the *Bauhinia* homeolog 2; and each *Bauhinia* homoeolog has a K_s of 0.25 with respect to the other *Bauhinia* homoeolog. The relationship between the species history and the gene tree is complicated by the hypothesized slower substitution rate in Cercis.

Precedent for a significant period of species divergence prior to allopolyploidy is seen, for example, in *Arachis*: the allopolyploid *A. hypogaea* was formed, within about the last 10 thousand years, from the merger of *A. duranensis* and *A. ipaensis*, which diverged an estimated 2.16 Mya (Bertioli et al., 2016). Another similar example is in cotton, where the allotetraploid *Gossypium hirsutum* L. is a merger of genomes from progenitor species similar to the extant diploid species *G. ramondii* Ulbrich and *G. herbaceum* L. (Wendel, 1989; Flagel et al., 2012; Paterson et al., 2012) In this case, the diploid species diverged c. 5–10 Mya and merged to form *G. hirsutum* c. 1–2 Mya (Wendel, 1989; Fang et al., 2017).

The genus *Cercis* contains 10 species and all phylogenetic analyses to date have supported the genus as monophyletic. This is a well-defined group of north temperate trees (North America, Eurasia and eastern Asia). All species for which counts are available are diploid[2]. There appears to be relatively low genetic diversity within the genus based on plastid and nuclear ribosomal ITS sequences (Davis et al., 2002; Coskun and Parks, 2009). *C. chingii* ($n = 14$) is resolved as sister to the other species in the genus in the studies by Davis et al. (2002), and differs from the other species by its coriaceous, unwinged, dehiscent fruit. The other species are morphologically quite similar. It's not clear if one of the present day *Cercis* species could better represent an ancestral parental genome resulting in the whole genome duplication.

[2]http://www.tropicos.org/Project/IPCN

Cercis genes do appear to have evolved remarkably slowly (at least in the sense of accumulating point mutations that affect K_s and branch lengths). A tree calculated by algebraically solving evolutionary "distance paths" along a gene tree (**Figures 1**, **2**, lower right), using K_s-based branch lengths, shows a *Cercis* evolutionary rate less than a quarter that of *Bauhinia*, and roughly a tenth that of *Phaseolus* since the papilionoid WGD. The slow *Cercis* rate is also evident in many gene family trees, such as the two shown in **Figure 3**. The *matK* gene tree also shows remarkably short branches for *Cercis*. It is conceivable that the slower evolutionary rate seen in *Cercis* than other legumes might be partly due to the lack of WGD-derived "extra" genes in *Cercis* –perhaps presenting extra evolutionary constraints than for duplicated genes. The outcrossing, long-lived tree form might also constrain evolutionary rates (injecting older gametes into new progeny) – although of course these conditions are shared with many species.

CONCLUSION

The evidence from diverse sources indicates that *Cercis* may be unique among legume lineages in lacking any evidence for a WGD; that its last duplication event was probably the eudicot "gamma" triplication event; that the genomes of other Cercidoideae and all other legume subfamilies are likely to have been shaped by independent WGD events; that the most likely model for WGD and speciation timing in the Cercidoideae is allopolyploidy – with a *Cercis* progenitor contributing one subgenome to the allopolyploid *Bauhinia* progenitor; and lastly, that *Cercis* has evolved at a strikingly low rate since its divergence from other Cercidoideae. Taken together, these findings suggest that *Cercis* may serve as a useful genomic model for the legumes, likely representing the duplication status of the progenitor of all legumes.

AUTHOR CONTRIBUTIONS

SC, JD, and DF-B conceptualized the research. JS, AY, SC, and DF-B planned and constructed gene families. JS and AY conducted phylogenomic tests. AB and CS generated and assembled *CYCLOIDEA* sequences for species phylogenetic analyses. SC, JS, and AY drafted the manuscript. All authors reviewed and contributed to the manuscript.

FUNDING

This research was funded by the NSF project "Federated Plant Database Initiative for the Legumes" (#1444806), and the USDA Agricultural Research Service project 5030-21000-062-00D. The USDA is an equal opportunity provider and employer.

ACKNOWLEDGMENTS

We thank many legume systematists from the Legume Phylogenetic Working Group for assembling a large collection of *matK* sequences that were essential for placing our analysis in the context of the legume phylogenetic history; Nathan Weeks for providing infrastructure for computational biology work; and Jean-Michel Ane and Matthew Crook for providing access to genome assemblies and annotations for *Cercis*, *Chamaecrista*, *Mimosa*, and *Nissolia*.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.00345/full#supplementary-material

TABLE S1 | Histograms of K_s values for selected species pairs, calculated using top BLAST matches and calculated using gene pairs from genomic synteny features.

TABLE S2 | Counts of species in legume gene families.

TABLE S3 | Sequence IDs and accessions for *CYCLOIDEA*-like genes from species in the Cercidoideae.

TABLE S4 | Sequences for legume matK genes, and chromosome counts for legume species, with modal chromosome counts per genus.

TABLE S5 | Legume genome size estimates.

DATA SHEET S1 | legume.genefam.esm.RDQM.SD01_65_hmmalign Full alignments of all legume gene families that contain *Cercis* and/or *Bauhinia*.

DATA SHEET S2 | legume.genefam.esm.RDQM.SD02_65_hmmalign_trim2 Alignments of all legume gene families that contain *Cercis* and/or *Bauhinia*, trimmed to HMM "match states," prior to calculation of phylogenetic trees.

DATA SHEET S3 | SD03_70_trees_combined Phylogenetic trees, calculated using RAxML, and rooted by the nearest non-legume outgroup species.

DATA SHEET S4 | SD04_71_trees_reduced_trees Phylogenetic trees as in SD03 but with same-species terminal branches reduced to a single sequence representative, for terminals with branch lengths < 0.2.

DATA SHEET S5 | SD05_tree_legcyc_and_outgrps7boot_color.nh.txt Phylogenetic tree, calculated using RAxML, for *CYCLOIDEA*-like genes from species in the Cercidoideae.

DATA SHEET S6 | Phylogenetic tree image for all legume genera with chromosome counts, with modal chromosome counts per genus, colored by subfamily.

DATA SHEET S7 | SD11_LPWG_chrom_counts_by_genus3.tree.txt Phylogenetic tree file, in phylip format, corresponding with the image in SD10.

DATA SHEET S8 | Synteny plots between *Cercis canadensis*, *Phaseolus vulgaris*, and *Prunus persica*.

DATA SHEET S9 | SD09_quota_tables Results from quota-alignment pipeline for comparisons between *C. canadensis*, *Phaseolus vulgaris*, and *Prunus persica*.

REFERENCES

Angiosperm Phylogeny Group, Byng, J. W., Chase, M. W., Christenhusz, M. J. M., Fay, M. F., 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

Arumuganathan, K., and Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* 9, 208–218. doi: 10.1007/BF02672069

Battenberg, K., Potter, D., Tabuloc, C. A., Chiu, J. C., and Berry, A. M. (2018). Comparative Transcriptomic analysis of two actinorhizal plants and the legume medicagotruncatula supports the homology of root nodule symbioses and is congruent with a two-step process of evolution in the nitrogen-fixing clade of angiosperms. *Front. Plant Sci.* 9:1256. doi: 10.3389/fpls.2018.01256

Bennett, M. D., and Leitch, I. J. (2005). Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Ann. Bot.* 95, 45–90. doi: 10.1093/aob/mci003

Bennett, M. D., and Leitch, I. J. (2011). Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Ann. Bot.* 107, 467–590. doi: 10.1093/aob/mcq258

Bennett, M. D., and Leitch, I. J. (2012). *). Plant C-Values Database (Release 6.0, Dec. 2012)*. Available: http://data.kew.org/cvalues/ [accessed April 10, 2014].

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The arabidopsis information resource: making and mining the 'Gold Standard' annotated reference plant genome. *Genesis* 53, 474–485. doi: 10.1002/dvg.22877

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., et al. (2016). The genome sequences of arachis duranensis and arachis ipaensis, the diploid ancestors of cultivated peanut. *Nat. Genet.* 48, 438–446. doi: 10.1038/ng.3517

Brottier, L., Chaintreuil, C., Simion, P., Scornavacca, C., Rivallan, R., Mournet, P., et al. (2018). A phylogenetic framework of the legume genus aeschynomene for comparative genetic analysis of the Nod-dependent and Nod-independent symbioses. *BMC Plant Biol.* 18:333. doi: 10.1186/s12870-018-1567-z

Bruneau, A., Mercure, M., Lewis, G. P., and Herendeen, P. S. (2008). Phylogenetic patterns and diversification in the caesalpinioid legumes this paper is one of a selection of papers published in the special issue on systematics research. *Botany* 86, 697–718. doi: 10.1139/B08-058

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Cannon, S. B., McKain, M. R., Harkess, A., Nelson, M. N., Dash, S., Deyholos, M. K., et al. (2015). Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Mol. Biol. Evol.* 32, 193–210. doi: 10.1093/molbev/msu296

Cardoso, D., de Queiroz, L. P., Pennington, R. T., de Lima, H. C., Fonty, É, Wojciechowski, M. F., et al. (2012). Revisiting the phylogeny of papilionoid legumes: new insights from comprehensively sampled early-branching lineages. *Am. J. Bot.* 99, 1991–2013. doi: 10.3732/ajb.1200380

Citerne, H. L., Luo, D., Pennington, R. T., Coen, E., and Cronk, Q. C. B. (2003). A phylogenomic investigation of CYCLOIDEA-Like TCP genes in the leguminosae. *Plant Physiol.* 131, 1042–1053. doi: 10.1104/pp.102.016311

Citerne, H. L., Pennington, R. T., and Cronk, Q. C. B. (2006). An apparent reversal in floral symmetry in the legume Cadia is a homeotic transformation. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12017–12020. doi: 10.1073/pnas.0600986103

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Coskun, F., and Parks, C. R. (2009). A molecular phylogenetic study of red buds (Cercis L., fabaceae) based on its nrDNA sequences. *Pak. J. Bot.* 41, 1577–1586.

Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16, 738–749. doi: 10.1101/gr.4825606

Davis, C. C., Fritsch, P. W., Li, J., and Donoghue, M. J. (2002). Phylogeny and biogeography of Cercis (Fabaceae): evidence from nuclear ribosomal ITS and chloroplast ndhF sequences. *Syst. Bot.* 27, 289–302.

Dolezel, J., Bartoš, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry* 51A, 127–128. doi: 10.1002/cyto.a.10013

Doyle, J. J. (2012). "Polyploidy in Legumes," in *Polyploidy and Genome Evolution*, eds P. S. Soltis and D. E. Soltis (Heidelberg: Springer), 147–180. doi: 10.1007/978-3-642-31442-1-9

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., et al. (2017). Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* 49, 1089–1098. doi: 10.1038/ng.3887

Felsenstein, J. (1980). *PHYLIP*. Available at: http://evolution.genetics.washington.edu/phylip.html

Flagel, L. E., Wendel, J. F., and Udall, J. A. (2012). Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. *BMC Genomics* 13:302. doi: 10.1186/1471-2164-13-302

Goldblatt, P. (1981). Chromosome numbers in legumes II. *Ann. Mo. Bot. Gard.* 68, 551–557. doi: 10.2307/2398889

Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M. B., et al. (2018). Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* 361:eaat1743. doi: 10.1126/science.aat1743

Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E., and Lyons, E. (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* 33, 2197–2198. doi: 10.1093/bioinformatics/btx144

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldón, T. (2007). The human phylome. *Genome Biol.* 8:R109. doi: 10.1186/gb-2007-8-6-r109

Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python environment for tree exploration. *BMC Bioinformatics* 11:24. doi: 10.1186/1471-2105-11-24

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi: 10.1093/molbev/msw046

International Peach Genome Initiative[IPGI] (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* 45, 487–494. doi: 10.1038/ng.2586

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., McKain, M. R., McNeal, J., et al. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13:R3. doi: 10.1186/gb-2012-13-1-r3

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B.-K., et al. (2014). Genome sequence of mungbean and insights into evolution within Vigna species. *Nat. Commun.* 5:5443. doi: 10.1038/ncomms6443

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Lavin, M., Herendeen, P. S., and Wojciechowski, M. F. (2005). Evolutionary rates analysis of leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* 54, 575–594. doi: 10.1080/10635150590947131

Legume Phylogeny Working Group, Azani, N., Babineau, M., Bailey, C. D., Banks, H., Barbosa, A., et al. (2017). A new subfamily classification of the leguminosae based on a taxonomically comprehensive phylogeny – the legume phylogeny working group (LPWG). *Taxon* 66, 44–77. doi: 10.12705/661.3

Leitch, I. J., Hanson, L., Lim, K. Y., Kovarik, A., Chase, M. W., Clarkson, J. J., et al. (2008). The ups and downs of genome size evolution in polyploid species of Nicotiana (Solanaceae). *Ann. Bot.* 101, 805–814. doi: 10.1093/aob/mcm326

Lewis, G. P., Schrire, B., Mackinder, B., and Lock, M. (eds.) (2005). *Legumes of the World*, 1st Edn. Richmond: The Royal Botanic Gardens,Kew.

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. doi: 10.1126/science.290.5494.1151

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41, e121–e121. doi: 10.1093/nar/gkt263

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426. doi: 10.1093/oxfordjournals.molbev.a040410

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

Phytozome 12 (2018). *Cucumis Sativus v1.0 (Cucumber)*. Available at: https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Csativus [accessed December 14, 2018].

Ren, L., Huang, W., and Cannon, S. B. (2019). Reconstruction of ancestral genome reveals chromosome evolution history for selected legume species. *New Phytol.* (in press). doi: 10.1111/nph.15770

Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N. M., Salman-Minkov, A., et al. (2015). The chromosome counts database (CCDB) – a community resource of plant chromosome numbers. *New Phytol.* 206, 19–26. doi: 10.1111/nph.13191

Roberts, D. J., and Werner, D. J. (2016). Genome size and ploidy levels of cercis (Redbud) species, cultivars, and botanical varieties. *HortScience* 51, 330–333. doi: 10.21273/HORTSCI.51.4.330

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., et al. (2008). Genome structure of the legume, lotus japonicus. *DNA Res.* 15, 227–239. doi: 10.1093/dnares/dsn008

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* 46, 707–713. doi: 10.1038/ng.3008

Sinou, C., Forest, F., Lewis, G. P., and Bruneau, A. (2009). The genus Bauhinia s.l. *(Leguminosae)*: a phylogeny based on the plastid trn L– trn F region. *Botany* 87, 947–960. doi: 10.1139/B09-065

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* 57, 758–771. doi: 10.1080/10635150802429642

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

Tang, H., and Chapman, B. A. (2018). *My Collection of Light Bioinformatics Analysis Pipelines for Specific Tasks: Tanghaibao/bio-pipeline*. Available at: https://github.com/tanghaibao/bio-pipeline [accessed November 15, 2018].

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., et al. (2014). An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* 15:312. doi: 10.1186/1471-2164-15-312

Tang, H., Lyons, E., Pedersen, B., Schnable, J. C., Paterson, A. H., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* 12:102. doi: 10.1186/1471-2105-12-102

The International Peach Genome Initiative, Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., et al. (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* 45, 487–494. doi: 10.1038/ng.2586

The Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. doi: 10.1038/nature11119

Town, C. D., Cheung, F., Maiti, R., Crabtree, J., Haas, B. J., Wortman, J. R., et al. (2006). Comparative Genomics of brassica oleracea and arabidopsis thaliana reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* 18, 1348–1359. doi: 10.1105/tpc.106.041665

van Velzen, R., Holmer, R., Bu, F., Rutten, L., Zeijl, A., van Liu, W., et al. (2018). Comparative genomics of the nonlegume parasponia reveals insights into evolution of nitrogen-fixing rhizobium symbioses. *Proc. Natl. Acad. Sci. U.S.A.* 115, E4700–E4709. doi: 10.1073/pnas.1721395115

Varshney, R. K., Chen, W., Li, Y., Bharti, A. K., Saxena, R. K., Schlueter, J. A., et al. (2012). Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* 30, 83–89. doi: 10.1038/nbt.2022

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., et al. (2013). Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat. Biotechnol.* 31, 240–246. doi: 10.1038/nbt.2491

Wang, Y.-H., Wicke, S., Wang, H., Jin, J.-J., Chen, S.-Y., Zhang, S.-D., et al. (2018). Plastid genome evolution in the early-diverging legume subfamily cercidoideae (Fabaceae). *Front. Plant Sci.* 9:138. doi: 10.3389/fpls.2018.00138

Wendel, J. F. (1989). New world tetraploid cottons contain old world cytoplasm. *Proc. Natl. Acad. Sci. U.S.A.* 86, 4132–4136. doi: 10.1073/pnas.86.11.4132

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

Young, N. D., Debellé, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K., et al. (2011). The medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* 480, 520–524. doi: 10.1038/nature10625

Zhao, Z., Hu, J., Chen, S., Luo, Z., Luo, D., Wen, J., et al. (2019). Evolution of CYCLOIDEA-like genes in fabales: insights into duplication patterns and the control of floral symmetry. *Mol. Phylogenet. Evol.* 132, 81–89. doi: 10.1016/j.ympev.2018.11.007

Conflict of Interest Statement: 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: 12 July 2019
doi: 10.3389/fpls.2019.00864

Targeted Capture of Hundreds of Nuclear Genes Unravels Phylogenetic Relationships of the Diverse Neotropical Palm Tribe Geonomateae

Oriane Loiseau[1], Ingrid Olivares[2,3], Margot Paris[4], Marylaure de La Harpe[5], Anna Weigand[2], Darina Koubínová[1,6], Jonathan Rolland[1,7], Christine D. Bacon[8,9], Henrik Balslev[10], Finn Borchsenius[11], Angela Cano[12], Thomas L. P. Couvreur[13], César Delnatte[14], Frédérique Fardin[15], Marc Gayot[16], Fabian Mejía[17], Talita Mota-Machado[18], Mathieu Perret[19], Julissa Roncal[20], Maria José Sanin[17], Fred Stauffer[19], Christian Lexer[5], Michael Kessler[2†] and Nicolas Salamin[1†]*

[1] Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, [2] Department for Systematic and Evolutionary Botany, University of Zurich, Zurich, Switzerland, [3] Centre for Biodiversity and Environment Research, University College London, London, United Kingdom, [4] Department of Biology, Unit Ecology and Evolution, University of Fribourg, Fribourg, Switzerland, [5] Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria, [6] Natural History Museum of Geneva, Geneva, Switzerland, [7] Department of Zoology, University of British Columbia, Vancouver, BC, Canada, [8] Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden, [9] Gothenburg Global Biodiversity Centre, Gothenburg, Sweden, [10] Department of Bioscience, Biodiversity and Ecoinformatics, Aarhus University, Aarhus, Denmark, [11] Science Museums, Aarhus University, Aarhus, Denmark, [12] Cambridge University Botanic Garden, Cambridge, United Kingdom, [13] IRD, DIADE, University of Montpellier, Montpellier, France, [14] National Forestry Office, Fort-de-France, France, [15] Parc National de la Guadeloupe, Guadeloupe, France, [16] National Forestry Office, Guadeloupe, France, [17] Facultad de Ciencias y Biotecnología, Universidad CES, Medellin, Colombia, [18] Programa de Pós-Graduação em Biologia Vegetal, Departamento de Botânica, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [19] Department of Botany and Plant Biology, Conservatory and Botanical Garden of the City of Geneva, University of Geneva, Geneva, Switzerland, [20] Department of Biology, Memorial University of Newfoundland, St. John's, NL, Canada

The tribe Geonomateae is a widely distributed group of 103 species of Neotropical palms which contains six ecologically important understory or subcanopy genera. Although it has been the focus of many studies, our understanding of the evolutionary history of this group, and in particular of the taxonomically complex genus *Geonoma*, is far from complete due to a lack of molecular data. Specifically, the previous Sanger sequencing-based studies used a few informative characters and partial sampling. To overcome these limitations, we used a recently developed Arecaceae-specific target capture bait set to undertake a phylogenomic analysis of the tribe Geonomateae. We sequenced 3,988 genomic regions for 85% of the species of the tribe, including 84% of the species of the largest genus, *Geonoma*. Phylogenetic relationships were inferred using both concatenation and coalescent methods. Overall, our phylogenetic tree is highly supported and congruent with taxonomic delimitations although several morphological taxa were revealed to be non-monophyletic. It is the first time that such

OPEN ACCESS

Edited by:
Lisa Pokorny, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Spain

Reviewed by:
Karolina Heyduk, Yale University, United States
Zhen Li, Ghent University, Belgium

***Correspondence:**
Nicolas Salamin
nicolas.salamin@unil.ch

† Share co-last authorship

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

Received: *14 January 2019*
Accepted: *17 June 2019*
Published: *12 July 2019*

Citation:
Loiseau O, Olivares I, Paris M, de La Harpe M, Weigand A, Koubínová D, Rolland J, Bacon CD, Balslev H, Borchsenius F, Cano A, Couvreur TLP, Delnatte C, Fardin F, Gayot M, Mejía F, Mota-Machado T, Perret M, Roncal J, Sanin MJ, Stauffer F, Lexer C, Kessler M and Salamin N (2019) Targeted Capture of Hundreds of Nuclear Genes Unravels Phylogenetic Relationships of the Diverse Neotropical Palm Tribe Geonomateae. Front. Plant Sci. 10:864. doi: 10.3389/fpls.2019.00864

a large genomic dataset is provided for an entire tribe within the Arecaceae. Our study lays the groundwork not only for detailed macro- and micro-evolutionary studies within the group, but also sets a workflow for understanding other species complexes across the tree of life.

Keywords: Arecaceae, *Geonoma*, Neotropics, phylogenetic informativeness, phylogenomics, species complexes

INTRODUCTION

Palms (Arecaceae) are an important ecological component (Henderson, 2002; Couvreur et al., 2011) and a useful plant group of tropical ecosystems (Macia et al., 2011; Gruca et al., 2015). The palm family was recently advocated as a model group to understand the evolution of tropical rain forests (Baker and Couvreur, 2013) and numerous studies have investigated their phylogenetic relationships and systematics (Uhl and Dransfield, 1987; Baker et al., 1999; Dransfield et al., 2008; Baker and Dransfield, 2016). However, given the remarkably low rate of molecular evolution observed in palms (Wilson et al., 1990), phylogenetic studies at different taxonomic levels within the Arecaceae based on a few plastid or nuclear genes generally result in poorly resolved phylogenetic trees, especially at the species level (Roncal et al., 2005, 2008; Cuenca et al., 2008; Baker et al., 2011; Bacon et al., 2012a, 2016a,b, 2017; Meerow et al., 2015; Sanín et al., 2016).

The lack of informative genetic markers, combined with insufficient taxonomic sampling, currently limits our understanding of the phylogenetic relationships within the most diverse palm genera in the Neotropics, such as *Bactris* Jacq. ex Scop., *Chamaedorea* Willd., and *Geonoma* Willd. These three genera are mostly small shade-adapted palms and they contain the most abundant palm species in the understory of many Neotropical forests (Vormisto et al., 2004; Balslev et al., 2016, 2017; Ley-lopez and Avalos, 2017; Muscarella et al., 2018). They also often exhibit a high amount of intraspecific phenotypic variation (Roncal, 2006), which is hard to address with a taxonomic classification. This is exemplified by *Geonoma*, which, with 68 recognized species (Henderson, 2011), is the third most diverse palm genus in the Neotropics. *Geonoma* belongs to the tribe Geonomateae Luerss., together with five other genera. These five additional genera range in size from two (*Welfia* H. Wendl.) to 21 species (*Calyptrogyne* H. Wendl.), and are also small understory palms except for *Calyptronoma* Griseb. (three species) and *Welfia,* which can reach up to 15 m and 25 m, respectively. The tribe displays a wide geographical and ecological distribution, occurring from southern Mexico to south-eastern Brazil, including the Caribbean, with species growing from the lowlands up to 3,000 m elevation in the Andes. The tribe has been intensively studied and its main biological aspects, such as taxonomy (Wessels Boer, 1968; Zona, 1995; Stauffer et al., 2003; Henderson, 2005, 2011, 2012; Henderson and Villalba, 2013), ecology (Chazdon, 1992; Knudsen et al., 1998; Sampaio and Scariot, 2008; Pizo and Almeida-Neto, 2009), and phylogenetic relationships (Roncal et al., 2005, 2010, 2011, 2012) have been characterized to some extent. Considerable research has also been dedicated to investigate the phenotypically widely variable species complexes that represent 20% of the species of *Geonoma* (Borchsenius, 2002; Henderson and Martins, 2002; Roncal, 2006; Roncal et al., 2007; Henderson, 2011; Borchsenius et al., 2016). Despite all these efforts, the evolutionary history of *Geonoma* and Geonomateae remains only partially understood due to the paucity of DNA sequences, which so far are available only for three nuclear loci and approximately 60% of the species in the tribe.

Obtaining a robust phylogenetic hypothesis for the Geonomateae is therefore crucial to enable a reliable assessment of the systematic relationships of its lineages, but also to provide the foundation to assess the macroevolutionary patterns and the dynamics of diversification in this key palm group. The increasing affordability of next generation sequencing techniques, which offers the possibility to sequence hundreds of loci at a time, has already benefitted many plant phylogenetic studies (e.g., Nicholls et al., 2015; Sass et al., 2016; Mandel et al., 2017; Mitchell et al., 2017). For Arecaceae, while most of genome-scale data initially focused on commercially important species such as the oil palm (Uthaipaisanwong et al., 2012; Singh et al., 2013) and the date palm (Yang et al., 2010; Al-Mssallem et al., 2013), evolutionary biologists have put considerable effort in the last few years to generate genomic data across the whole family and are aiming at a species level phylogenetic tree of all palms (Comer et al., 2015, 2016; Heyduk et al., 2015; Barrett et al., 2016, 2018).

In this context, the recent development of several sequence capture kits for the Arecaceae (Heyduk et al., 2015; de La Harpe et al., 2019) represents an ideal opportunity to fill the gaps in palm phylogenomics. Here, using the bait kit developed by de La Harpe et al. (2019), we sequenced 4,184 genomic regions for 85% of the species of tribe Geonomateae, including 84% of the species of *Geonoma* and applied both standard and coalescent-based methods to reconstruct the phylogenetic relationships within the tribe. Using substantial intraspecific sampling, we assessed the validity of the species delimitations proposed by Henderson (2011) for the widespread and highly morphologically variable species complexes. We also estimated the phylogenetic informativeness of the DNA regions in the capture kit and proposed a smaller selection of the most useful genomic regions for phylogenetic studies at deep and shallow evolutionary scales within the Arecaceae. Our results show that these new molecular tools increase our understanding of the systematics and evolution in this important group of understory palms and open up new directions of research to test hypotheses about the factors underlying the diversification of species in palms.

MATERIALS AND METHODS

Taxon Sampling

We gathered a total of 312 samples of either silica-dried leaves or herbarium fragments from specimens stored at the herbarium of Geneva (G) and the Herbario Nacional Colombiano (COL), including 240 samples representing 57 (84%) of the 68 currently recognized species of *Geonoma* (**Supplementary Table S1**). Among the 11 missing species of *Geonoma*, eight are narrow endemics known only from the type collections (*G. deneversii* A. J. Hend., *G. dindoensis* A. J. Hend., *G. gentryi* A. J. Hend. and *G. operculata* A. J. Hend.) or less than ten herbarium specimens (*G. peruviana* A. J. Hend., *G. sanmartinensis* A. J. Hend., *G. schizocarpa* A. J. Hend. and *G. venosa* A. J. Hend.). Whenever possible, we sampled several individuals per species and included different subspecies. For widely distributed species, sample selection was designed to cover the greatest possible extant of their geographic distribution. Our sampling also included 65 individuals representing 25 species from the other five genera of the tribe Geonomateae (100% taxon sampling for *Asterogyne* H. Wendl, 61% for *Calyptrogyne*, 100% for *Calyptronoma*, 75% for *Pholidostachys* H. Wendl. Ex Hook. f., and 50% for *Welfia*), covering in total 85% of the tribe's species richness. For the purpose of computing the phylogenetic informativeness of the targeted genomic regions across the whole Arecaceae, we also included seven samples from phylogenetically more distant palm genera, belonging to subfamilies Arecoideae Burnett (*Bactris, Cocos* L., *Socratea* H. Karst, and *Wettinia* Poepp.), Ceroxyloideae Drude (*Ceroxylon* Bonpl. ex DC.), and Coryphoideae Burnett (*Licuala* Wurmb).

DNA Extraction, Dual-Indexed Library Preparation, and Target Capture Sequencing

DNA was extracted using the DNeasy® plant mini kit (Qiagen, Venlo, Netherlands) following the supplier's instructions. DNA quality and degradation were evaluated with agarose gels and a Nanodrop™TM spectrophotometer ND-1000 (Thermo Fisher Scientific, Waltham, MA, United States) and DNA was quantified with a Qubit® Fluorometer v 2.2 (Thermo Fisher Scientific, Waltham, MA, United StatesUSA). When possible, a total of 500 ng of DNA were used per sample for library preparation.

DNA samples were fragmented to 400 bp fragments with a bioruptor® ultrasonicator UCD-200TM-EX (Diagenode, Liège, Belgium) with six cycles of 30 s ON, and 90 s OFF. This step was omitted for samples with degraded DNA. Library preparations were performed following de La Harpe et al. (2019). Briefly, sample cleaning, end repair and A-tailing steps were carried out with a KAPA LTP library preparation kit (Roche, Basel, Switzerland), and adaptor ligation and adaptor fill-in reactions steps (Meyer and Kircher, 2010).

A set of 60 dual-index primers were used for amplification, as recommended by Kircher et al. (2011), to avoid inaccuracies in multiplex sequencing. Two sets of 7 bp indexes were generated using the create_index_sequences.py Python program (Meyer and Kircher, 2010): one set of 30 indexes for the P5 Illumina primers, and one set of 30 indexes for the P7 Illumina primers. The index lists were chosen to contain a balanced subset of indexes with an edit distance of 4 to reduce the chance of conversion by sequencing and amplification errors. Adaptor and primer sequences are described in **Supplementary Table S2**. Eight cycles of PCR were used for most samples, except for 29 low quality and degraded samples for which 12 cycles of PCR were necessary to obtain sufficient DNA amount (**Supplementary Table S1**). Libraries were quantified with a Qubit® Fluorometer v 2.2. Target capture was performed using the custom kit PopcornPalm developed by de La Harpe et al. (2019) and deposited in Dryad[1]. This kit targets 4,051 genes and 133 non-genic putatively neutral regions. Target capture was conducted on pooled dual-indexed libraries following myBait® Custom Target Capture Kits protocol v3.0 (Arbor Biosciences, Ann Arbor, MI, United States), with 18 h incubation time at 65°C and 12 cycles of post-capture PCR reactions. Pools of 64 samples were used as template for each target capture hybridization reaction, using an initial amount of 1.2 μg of pooled libraries. The pooled target capture reactions were quantified with a Qubit® Fluorometer v 2.2 before sequencing with an Illumina HiSeq3000 sequencer in paired-end 2 × 150 bp mode.

Read Trimming, Mapping, and SNP Calling

Reads were first trimmed with the program condetri v2.2 (Smeds and Künstner, 2011) using a base quality score of 20 as high-quality threshold parameter before mapping to the *Geonoma undata* Klotzsch pseudoreference genome described in de La Harpe et al. (2019) with bowtie2 v2.2.5 (Langmead and Salzberg, 2012) and the very-sensitive-local option. Only reads that mapped at a unique location in the genome were kept for analysis.

Before variant calling, PCR duplicates were masked with the software Picard v1.119[2], and reads were realigned around indels and base-recalibrated using GATK v3.8 (McKenna et al., 2010). SNPs were then called for targeted genomic regions using UnifiedGenotyper of GATK v3.8 using the EMIT_ALL_SITES option in order to obtain the full sequence of the targets. The main advantage of paired-end 2 × 150 bp read sequencing is the potential recovery of adjacent regions to the exonic targets. For this reason, the entire sequence including UTRs, exons and introns was called for each gene. Sites were filtered with the following parameters using VCFtools v0.1.13 (Danecek et al., 2011): minimum quality >20, no indel allowed, minimum depth of 8× per sample, and maximum of 50% of missing data. For each genomic region the alignment in fasta format was generated using the program vcf-tab-to-fasta[3].

Selection of Most Informative Genomic Regions

Because the bait kit developed by de La Harpe et al. (2019) for micro- and macro-evolutionary analyses in palms is large

[1]https://doi.org/10.5061/dryad.3v9v238

[2]http://broadinstitute.github.io/picard

[3]https://github.com/JinfengChen/vcf-tab-to-fasta

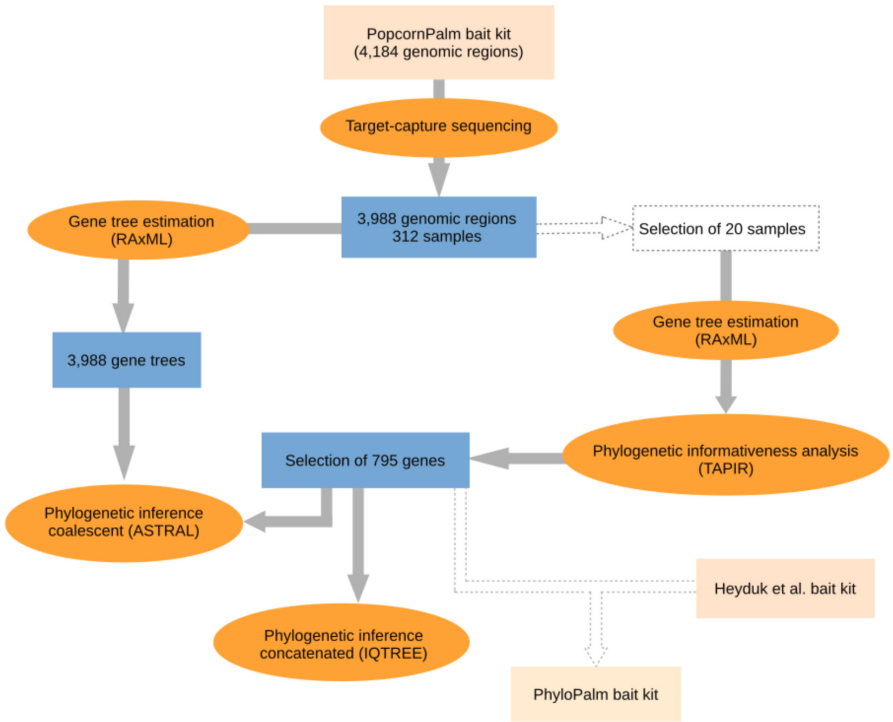

FIGURE 1 | Flowchart illustrating the steps of the phylogenomic analyses. The selection of 795 genes was combined with the bait kit from Heyduk et al. (2015) to provide a new bait kit for future phylogenomic studies.

(over 4,000 genomic regions) and contains several fast-evolving DNA regions that are not necessarily useful for phylogenomic studies, we selected a subsample of the most informative genomic regions which we then used to infer the species tree of the Geonomateae. Additionally, we made available a new bait kit for future phylogenomic studies in palms, which combines the subset of genes presented here with the genes from the Heyduk et al.'s kit (2015). Our workflow for gene selection and phylogenomic analyses is summarized in **Figure 1**. In order to maximize the phylogenetic informativeness of the retained genes for the Arecaceae and not only for tribe Geonomateae, the selection steps were performed on a dataset which contained species from three different Arecaceae subfamilies (Arecoideae, Ceroxyloideae, and Coryphoideae). First, we estimated the phylogenetic informativeness for each gene at different geological time intervals with the program TAPIR (Pond and Muse, 2005; Townsend, 2007; Faircloth et al., 2012). For each alignment, TAPIR estimates the site rates under the best-fitting substitution model and further computes a quantitative measure of the power of the gene to resolve the branching order at different depths of a given phylogenetic tree. To reduce computing time, the analysis was performed on a subset of 20 out of the 312 samples sequenced, which were selected to represent a wide range of evolutionary time scales, from intra-specific variability up to 88 Ma of divergence. The selection included three species of *Geonoma* (including four samples of *G. deversa*), two species of *Asterogyne*, two species of *Calyptrogyne*, as well as *Welfia regia* H. Wendl., *Bactris gasipaes* Kunth, *Cocos nucifera* L., *Socratea exorrhiza* (Mart.) H. Wendl, *Wettinia maynensis* Spruce, *Ceroxylon alpinum* Bonpl. ex DC., and two species of *Licuala*. Because TAPIR does not accept missing data, we only considered the genes for which sequence data were available for all 20 samples. Details for this analysis can be found in the **Supplementary Material**. Then, we selected the most appropriate genes for phylogenomic analyses according to the following criteria: (1) single-copy genes, (2) genes located on one of the 16 chromosomes of the *Elaeis guineensis* Jacq. reference genome (i.e., no gene on the extra low quality scaffolds), (3) genes absent from the bait kit of Heyduk et al. (2015) to avoid redundancy in the final bait set, (4) genes among the top 500 genes with the highest phylogenetic informativeness measure and/or with the highest mean bootstrap value per gene tree, (5) genes with a minimum mean bootstrap value per gene tree >60, and (6) genes with a minimum of five baits covering their exonic regions. We constrained our selection to a total of 17,091 baits to obtain a maximum of 20,000 baits when combined with the 2,909 baits of the Heyduk's kit (Heyduk et al., 2015). This option thus allows for coherence among different studies and maximizes the informativeness of the data at the lowest possible cost, since the smallest kit size available at the Arbor Biosciences company (Ann Arbor, MI, United States) is of 20,000 baits.

Phylogenetic Inference

Phylogenetic trees were estimated using both maximum likelihood and coalescent based methods. We used the software IQ-TREE (Nguyen et al., 2015) to estimate, under the maximum

likelihood criteria, the topology and branch lengths of the phylogenetic tree for all samples based on the concatenated analysis of the reduced set of genes satisfying the criteria described above. We partitioned the data by gene (Chernomor et al., 2016), using a GTR+GAMMA model of substitutions for each gene, and estimated support using the ultrafast bootstrap option (Hoang et al., 2018). We did not perform model testing because parameter rich models such as GTR+G and GTR+G+I have been shown in simulations to suffice for phylogeny reconstruction (Hoff et al., 2016; Abadi et al., 2019). The consensus tree obtained from this analysis was visualized using Figtree v1.4.3 (Rambaut, 2012). Next, we applied a coalescent approach which takes into account gene tree incongruence due to incomplete lineage sorting (Liu et al., 2009). We used ASTRAL 4.10.12, a two-step coalescent-based method that estimates the species tree given a set of gene trees (Mirarab et al., 2014; Mirarab and Warnow, 2015). All gene trees were first estimated with RaxML 8.2.10 (Stamatakis, 2014), using the GTR+GAMMA model of substitution and support estimated with the -autoMRE option. We then performed ASTRAL on the reduced data set and obtained a measure of branch support by computing local posterior probabilities. The impact of including weakly informative genes in two-step coalescent analyses is debated and while some studies showed that it can help to resolve difficult nodes (Blom et al., 2017) others argued that it reduces the accuracy of species tree estimation (Liu et al., 2015). To test whether including less informative genes would improve our phylogenetic inference we also applied ASTRAL to our full genomic dataset. Additionally, we computed quartet support values to measure the level of gene tree incongruence in our dataset and plotted quartet support for the three possible topologies at each branch using a python script[4]. All trees were rooted using the two *Licuala* species as outgroup.

RESULTS

Target Capture Sequencing

In total, we recovered DNA sequences for 3,988 genomic regions out of 4,184. On average, we obtained 2,064,810 reads per sample (**Supplementary Table S1**). After filtering, a total of 7,438,988 high quality bases including 2,288,308 SNPs were obtained with an average coverage of 30.8× and only 9.3% of missing data. When considering only the samples of *Geonoma*, 1,102,445 SNPs were recovered.

Phylogenetic Informativeness

Across our data set, phylogenetic informativeness increased with increasing evolutionary divergence times (**Figure 2**). After applying the selection step, the reduced dataset of 17,091 baits contained 795 genes, ranging from 1,108 to 12,710 bp in length. The corresponding bait kit combining our 795 genes with Heyduk's baits (Heyduk et al., 2015) is available at the Arbor Biosciences company (Ann Arbor, MI, United States) under the name "PhyloPalm."

Phylogenetic Inference

The total length of the concatenated alignment of the 795 selected genes was 3,064,021 bp. Phylogenetic trees obtained from the different datasets and methods had largely congruent topologies, except for the sister group of Clades XII-XIV (see section "Discussion" for clades numbers). This corresponded to Clade XI in the coalescent analysis of the 795 genes (with local posterior probability [LPP] of 0.59, **Figure 3**) and to Clades IX-X both in the concatenated analysis (with bootstrap support [BS] of 100%, **Figure 4**) and the coalescent analysis of the full dataset (with LPP of 0.66). For the 795 genes dataset, the support was slightly higher in the phylogenetic tree obtained with IQ-TREE (96% of nodes with BS >90, **Figure 4**) than with ASTRAL (89% of LPP >0.9, **Figure 3**). In the coalescent analyses, support increased with the size of the gene set, with 96% of branches having a LPP >0.9 in the phylogenetic tree obtained from the complete dataset of 3,988 genes. This is expected since the LPP are dependent on the discordance among gene trees but also the number of gene trees analyzed (Sayyari and Mirarab, 2016). Additionally, quartet support values indicated that gene tree incongruence is widespread across the phylogeny (**Figure 3**). In all analyses *Calyptronoma* was recovered paraphyletic, with *C. plumeriana* (Mart.) Lourteig and *C. rivalis* (O.F. Cook) L.H. Bailey more closely related to *Calyptrogyne* than to *C. occidentalis* (Sw.) H.E. Moore. The remaining five genera of tribe Geonomateae were recovered as monophyletic, with BS of 100% in the maximum likelihood phylogenetic tree and posterior probabilities of 1 in the coalescent phylogenetic trees.

DISCUSSION

The tribe Geonomateae is an ideal group to study plant evolutionary history in Neotropical rainforests for several reasons. First, it comprises the third largest genus of all Neotropical palms. Second, its species are distributed across all habitat types along the Andean and Central-American mountains as well as the Pacific, Caribbean and Amazonian lowlands, and in many of these areas they represent an important floristic element. Finally, *Geonoma* includes several species complexes with tremendous morphological variation which renders the taxonomic delimitation of species challenging. Because of these interesting characteristics, the systematics (Henderson et al., 1995; Henderson, 2011), ecology (Chazdon, 1991; Rodriguez-Buritica et al., 2005), and evolution (Roncal et al., 2011, 2012) of *Geonoma* have received significant attention. However, previous phylogenetic analyses relied on limited taxonomic and molecular sampling, thus preventing a detailed understanding of the phylogenetic relationships within the group. In particular, the phylogenetic trees recovered by Roncal et al. (2005, 2010, 2011, 2012) were not fully resolved and the status of species complexes had not been investigated. In this study, we addressed these shortcomings by applying

[4]https://github.com/sidonieB/scripts/blob/master/GetQpiechartsFromASTRAL.py

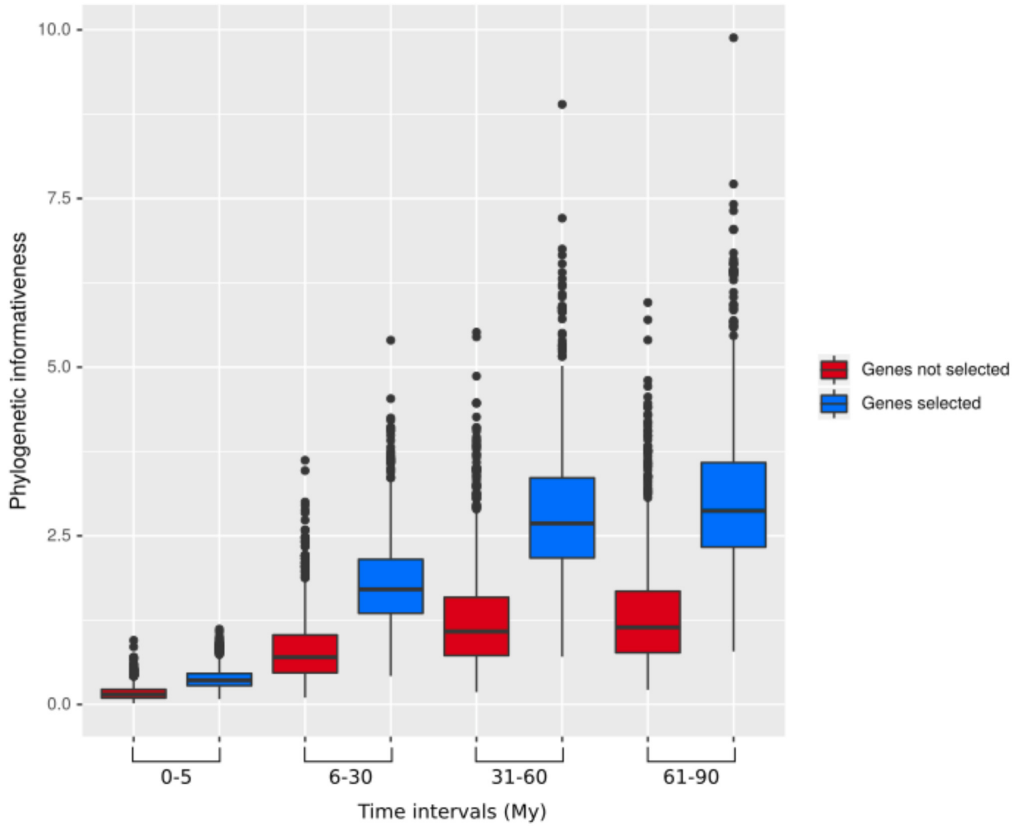

FIGURE 2 | Phylogenetic informativeness of the 795 selected genes (blue) and the remaining genes (red) over different evolutionary time intervals.

a target-capture approach, using the baits developed by de La Harpe et al. (2019) to sequence nearly 4,000 genes for 57 species of *Geonoma* and 25 species of the closely related genera of tribe Geonomateae. We performed concatenation and coalescent-based phylogenetic inferences which resulted in highly similar topologies, despite substantial amount of gene tree incongruence across the phylogeny. We showed that only a fraction of our complete genomic dataset was sufficient to resolve phylogenetic relationships within the Geonomateae (**Figures 3**, **4**).

Implications for the Systematics of Tribe Geonomateae

Phylogenetic relationships between the six genera of Geonomateae were so far poorly understood since various studies recovered different topologies (e.g., Baker et al., 2009; Roncal et al., 2010, 2011, 2012). Here, we were able to fully resolve the intergeneric relationships within the tribe but quartet support values in ASTRAL indicate a high level of gene tree incongruence (**Figure 3**), which problably explains the contrasted findings of the previous studies. We hypothesize that incomplete lineage sorting caused by the rapid divergence of the genera within tribe Geonomateae, as suggested by the very short branches in the maximum likelihood phylogenetic tree (**Figure 4**), is responsible for the observed gene tree discordance. Additionally, we confirmed the previously hypothesized paraphyly of the Caribbean endemic genus *Calyptronoma* and consequently advocate that it should be synonymized under *Calyptrogyne,* the sister group of *Geonoma.* The unique sample of *P. synanthera* (Mart.) H.E. Moore subspecies *synanthera* did not cluster with our two samples of *P. synanthera* subspecies *robusta* (Trail) A.J. Hend., but was instead recovered as sister to *P. sanluiensis* A. J. Hend. *Pholidostachys synanthera* subspecies *synanthera* and *P. sanluiensis* are Andean taxa which co-occur in the Central Cordillera of Colombia whereas *P. synanthera* subspecies *robusta* is a lowland West-Amazonian taxa. Additional sampling would be needed to test whether this placement is the result of hybridization between *P. synanthera* subspecies *synanthera* and *P. sanluiensis* or whether subspecies *synanthera* and *robusta* are actually separate species. Finally, we were able to assess the robustness of the current taxonomy for *Geonoma*, the largest and taxonomically most challenging genus of the tribe, in which the large degree of phenotypic variation has complicated species delimitations for a long time.

Phylogenetic Clades Within *Geonoma*

Based on our coalescent phylogeny and following the most recent phylogenetic reconstructions of the genus (Henderson, 2011; Roncal et al., 2011), we recognize 14 well-supported clades within

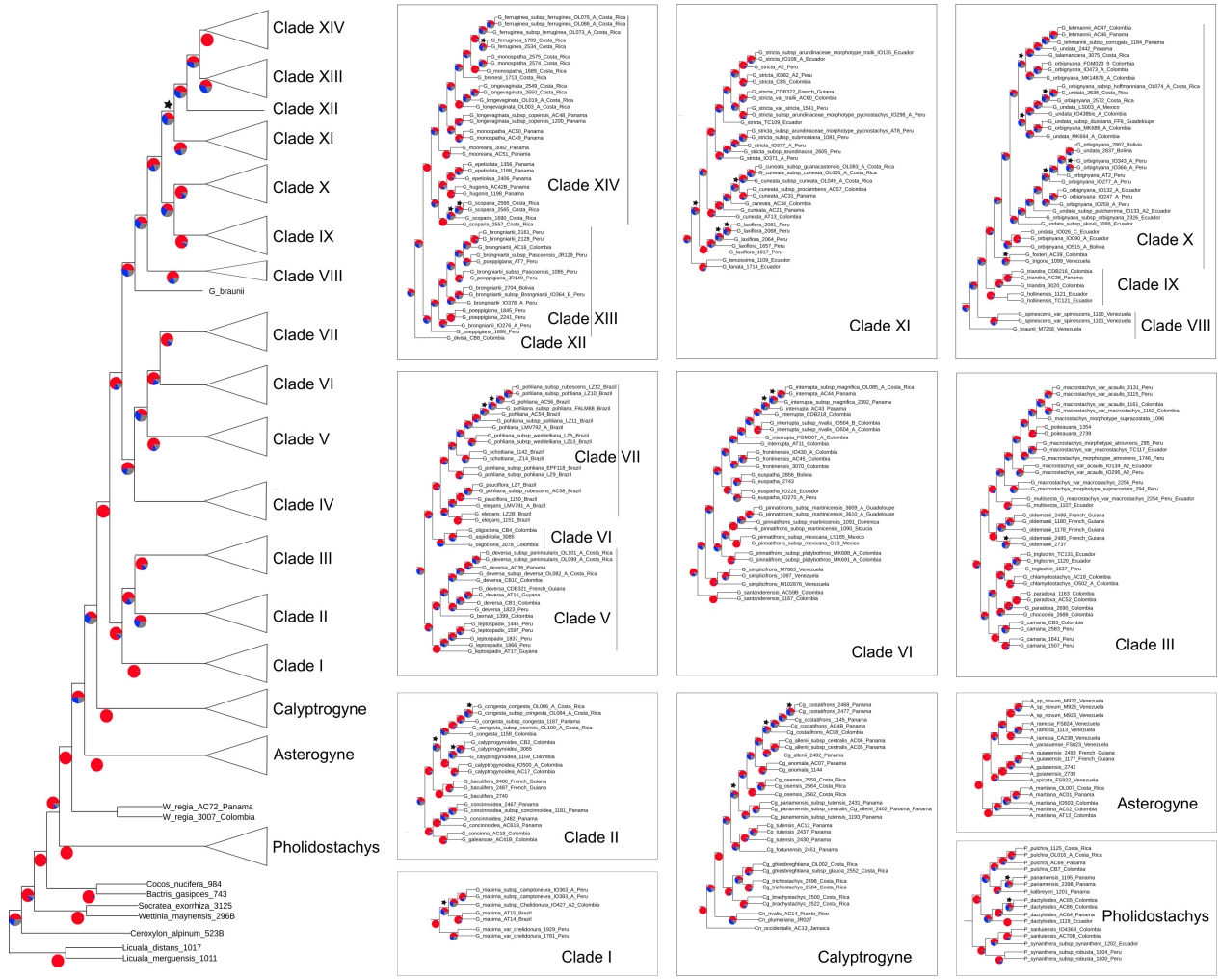

FIGURE 3 | Cladogram inferred with ASTRAL on the set of 795 gene trees. Pie charts indicate for each branch the percentage of gene trees aggreing with the topology of the species tree (red) and the percentage of gene trees supporting the other two alternative topologies (blue and gray). Stars indicate branches with LPP below 0.9.

Geonoma (**Figures 3**, **5**). We compare our findings with those from the maximum parsimony analysis of 30 morphological traits in the last revision of the group (Henderson, 2011). We use numbers to refer to clades to avoid confusion with the clade names previously used by Roncal et al. (2011) and Henderson (2011).

Clade I

This clade comprises a single variable species, *G. maxima* (Poit.) Kunth. It was included by Henderson (2011) in his *G. macrostachys* clade, which corresponds to Clade III in our analysis (**Figure 5**). *Geonoma maxima* differs, however, from all species of that clade in having a locular epidermis without an operculum, and a higher number of rachillae (4–50 vs. 1–9 rachillae; Henderson, 2011). Henderson (2011) recognized 11 subspecies within this primarily Amazonian lowland species, but the two subspecies with several individuals sampled (*G. maxima* subsp. *camptoneura* (Burret) A.J. Hend. and *G. maxima* subsp. *chelidonura* (Spruce) A.J. Hend) did not form monophyletic groups (**Figures 3**, **4**).

Clade II

This clade comprises six species (*G. baculifera* (Poit.) Kunth, *G. calyptrogynoidea* Burret, *G. concinna* Burret, *G. concinnoidea* A.J. Hend, *G. congesta* H, Wendl. ex Spruce, *G. galeanoae* A.J. Hend; **Figure 5**) that mostly grow in the lowlands of the Chocó region from Costa Rica to north-western Ecuador, with only *G. baculifera* occurring in north-eastern Amazonia and the Guianas. This clade corresponds to Henderson's (2011) *G. congesta* clade, which is characterized by the prophyll surfaces with close, equal, parallel, and non-dividing ridges. Henderson (2011) did not include *G. galeanoae* in this clade despite also sharing this trait, because its position in his morphology-based maximum parsimony tree was unresolved. Our analysis firmly recovers that species as a member of the clade and as sister to *G. concinna* with strong support (BS of 100, **Figure 4** and LPP

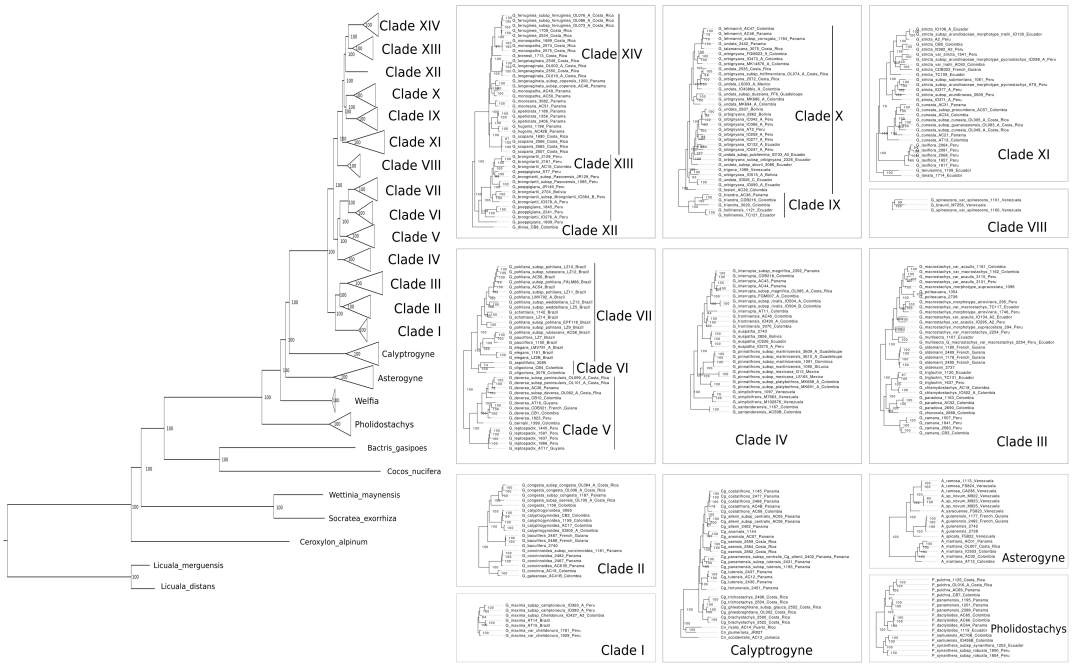

FIGURE 4 | Maximum likelihood phylogeny inferred from the concatenated alignment of the 795 selected genes. Numbers indicate BS.

of 1, **Figure 3**) emphasizing the taxonomic relevance of this character of the inflorescence bract. Henderson (2011) further recognized two subclades that were also recovered here with good support (BS of 100, **Figure 4** and LPP of 1, **Figure 3**): one including *G. concinna* and *G. concinnoidea* (and *G. galeanoae* in our study), and the other with the remaining species. This latter subclade is characterized by the non-homoplasious character state of staminodial tubes of non-fertilized pistillate flowers projecting and persistent after anthesis (Henderson, 2011).

Clade III

The third clade includes nine species [*G. camana* Trail, *G. chlamydostachys* Galeano, *G. chococola* Wess. Boer, *G. macrostachys* Mart., *G. multisecta* (Burret) Burret, *G. oldemannii* Granv., *G. paradoxa* Burret, *G. poiteauana* Kunth, *G. triglochin* Burret; **Figure 5**] from the Amazonian lowlands and adjacent regions, except for *G. paradoxa* from the Pacific coast of Colombia and Ecuador. It largely corresponds to Henderson's (2011) *G. macrostachys* clade except for three species (*G. deneversi*, *G. schizocarpa*, and *G. umbraculiformis*) that we did not sequence and *G. maxima*, which we recovered as an independent clade (see above). We recovered two subclades, one composed of *G. macrostachys*, *G. multisecta*, *G. poiteauana*, and *G. oldemannii* and the other of *G. camana, G. chlamydostachys*, *G. chococola, G. paradoxa,* and *G. triglochin* (each clade with bootstrap value of 100, **Figure 4** and LPP of 1, **Figure 3**). These clades did not correspond to the two subgroups recognized by Henderson (2011) in his *G. macrostachys* clade. *Geonoma macrostachys* is one of the morphologically and taxonomically most complex species in the genus, and Henderson (2011) divided it into several morphotypes, some of which behave as sympatric taxa at the local scale (Roncal, 2006; Roncal et al., 2007; Borchsenius et al., 2016). However, the 12 individuals of *G. macrostachys* from four morphotypes that we sampled show that, at larger geographic scales, morphotypes do not cluster into monophyletic groups. The species *G. poiteauna*, which used to be treated as a variety of *G. macrostachys* (Henderson et al., 1995) and subsequently raised at the species level (Henderson, 2011), is recovered here as nested within *G. macrostachys* (**Figures 3**, **4**).

Clade IV

This clade includes six species (*G. euspatha* Burret, *G. frontinensis* Burret, *G. interrupta* (Ruiz & Pav.) Mart, *G. pinnatifrons* Willd, *G. santanderensis* Galeano & R. Bernal, *G. simplicifrons* Willd; **Figure 5**) that largely occur on lower mountain slopes from Costa Rica to Bolivia and northeastern Brazil, as well as in the Antilles. It is essentially identical to Henderson's (2011) *G. interrupta* clade, except for *G. santanderensis* here recovered as sister to the other five species with strong support (BS of 100, **Figure 4** and LPP of 1, **Figure 3**) whereas it was placed within the *G. stricta* clade in Henderson's maximum parsimony analysis. Although *G. santanderensis* shares several specific morphological traits with *G. aspidiifolia* Spruce and *G. oligoclona* Trail (such as internodes covered with reddish or brownish scales, rachillae surfaces with spiky, fibrous projections or ridges, staminodial tubes lobed at the apex with the lobes not spreading at anthesis and not acuminate) our phylogenetic analyses reveal that these are homoplasic characters since the species are not closely related. Also, of the two subspecies of *G. interrupta* we sampled, the monophyletic subspecies *rivalis*, endemic of the Central Cordillera in Colombia, is nested within the geographically widespread subspecies *maxima*. The latter is also phenotypically

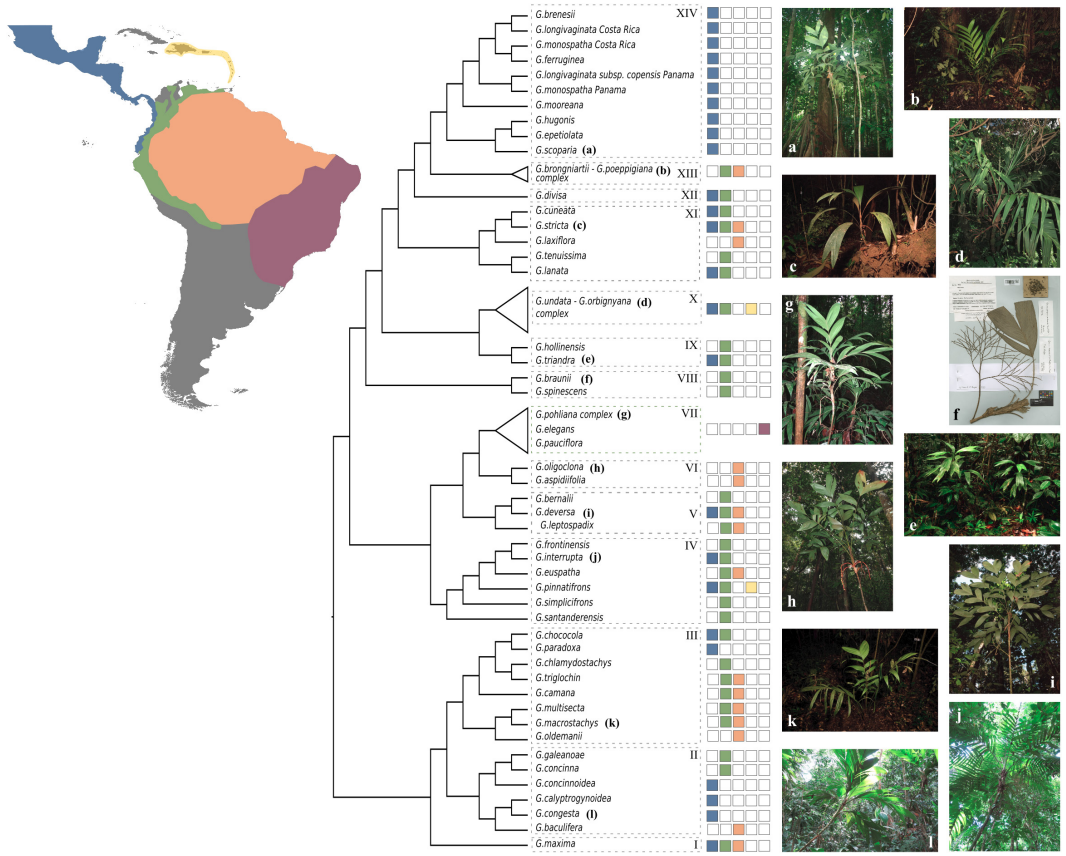

FIGURE 5 | Summary cladogram of the phylogenetic relationships in *Geonoma*. Dashed boxes with numbers refer to clades mentioned in the discussion. Colored squares show the geographic distribution of species. Photos by Ingrid Olivares **(a)**, FP7-PALMS project archive **(b,c,h,i,k)**, Rodrigo Cámara Leret **(e)**, Fred Stauffer **(f)**, Oriane Loiseau **(d,j,l)**, and Talita Mota Machado **(g)**.

more variable but never includes the rheophytic leaf morphology of subspecies *rivalis*.

Clade V

This clade includes three species (*G. bernalii* A. J. Hend, *G. deversa* (Poit.) Kunth, *G. leptospadix* Trail; **Figure 5**) occuring from Costa Rica to Peru and the Guianas. In Henderson's (2011) maximum parsimony tree, *G. bernalii* belonged to the *G. lanata* clade, whereas the other two species were placed in an unresolved polytomy. *Geonoma deversa* and *G. leptospadix* are variable and very widespread lowland species which probably hybridize in northeastern Brazil and the Guianas as suggested by the observation of specimens with intermediate morphology (Henderson, 2011) but none of the putative hybrid was sampled here. *Geonoma bernalii* occurs in northern Colombia and was previously identified as *G. leptospadix*. In our phylogenetic tree, *G. leptospadix* appears as sister to a group formed by *G. bernalii* and *G. deversa*.

Clade VI

This clade includes two morphologically very similar species (*G. aspidiifolia* and *G. oligoclona*; **Figure 5**) from Amazonia and the Guianan highlands. In fact, out of the two specimens of *G. oligoclona*, one is recovered more closely related to the single specimen of *G. aspidiifolia* (with BS of 74, **Figure 4** and LPP of 1, **Figure 3**). In the absence of additional individuals of *G. aspidiifolia*, it is premature to conclude whether they actually represent a single variable species or two closely related species. In Henderson's (2011) maximum parsimony tree these two species were placed within the *G. stricta* clade and were closely related to *G. santanderensis*.

Clade VII

This clade includes four species (*G. elegans* Mart., *G. pauciflora* Mart., *G. pohliana* Mart., and *G. schottiana* Mart.; **Figure 5**) from the Brazilian Atlantic Forest and the Cerrado. It corresponds to Henderson's (2011) *G. schottiana* clade and to Roncal et al. (2011) Brazilian Cerrado + Mata Atlantica clade. It appears that this small radiation of south-eastern Brazilian species resulted from a single colonization event that gave rise to four species with considerable inter- and intraspecific variation. In our phylogeny, the two individuals of *G. schottiana* were recovered as nested within *G. pohliana*, indicating that similarly to what is observed in other species complexes, morphological taxa are not always underpinned by strong genetic differentiation. *Geonoma pohliana* is an extremely variable species complex

which Henderson (2011) subdivided into 11 subspecies. The three subspecies sampled in our analysis appeared to be randomly mixed. The other two Brazilian species *G. elegans* and *G. pauciflora*, were recovered paraphyletic (**Figures 3**, **4**).

Clade VIII

This clade includes two morphologically similar endemic species from Venezuela (*G. spinescens* H. Wendl. ex Burret and *G. braunii* (Stauffer) A.J. Hend; **Figure 5**). Little DNA was obtained from the three herbarium samples and in fact *G. braunii* was recovered as sister taxa to clades VIII-XVI in the two ASTRAL analyses (**Figure 3**). However, we believe that this is caused by the lack of DNA sequences for *G. braunii.* Therefore, despite this uncertainty, we decided to follow the topology of the concatenated analysis and treat the two species as part of a single clade because it is coherent with the fact that *G. braunii* used to be considered a variety of *G. spinescens* (Stauffer, 1997). Henderson (2011) treated them as distinct species based on the flower pits alternately arranged in the former and spirally arranged in the latter. The two species were included within Henderson's (2011) *G. lanata* clade. In our concatenated analysis, *G. spinescens* appears paraphyletic because one of the two samples of *G. spinescens* was found sister to *G. braunii* (with bootstrap support of 99, **Figure 4**) but given the low amount of sequence data for *G. braunii* we can not advocate either of the two possible taxonomic treatment.

Clade IX

This clade includes two species (*G. hollinensis* A.J. Hend, Borchs & Balslev, and *G. triandra* (Burret) Wess. Boer; **Figure 5**) that are distributed from Panama to Ecuador and occur at similar elevations. The geographic distribution of these sister species (*G. hollinensis* restricted to north-eastern Ecuador and *G. triandra* found from north-western Ecuador to southern Panama) suggest that vicariance was involved in their divergence. Both species have staminate flowers with three stamens, and were segregated as subgenus *Kalbreyera* by Wessels Boer (1968). Henderson (2011) placed them together with *G. occidentalis* in the *G. triandra* clade but we cannot confirm this placement since *G. occidentalis* was not included in our study.

Clade X

This clade includes samples of five species (*G. lehmannii* Dammer ex Burret, *G. orbignyana* Mart., *G. talamancana* Grayum, *G. trigona* (Ruiz & Pav.) A.H. Gentry and *G. undata*; **Figure 5**) that occur at high elevations from Mexico to Bolivia, also reaching the Lesser Antilles, plus *G. fosteri* A.J. Hend. It is largely congruent with the *G. undata* clade of Henderson (2011) and the Andes + Central American Mountains clade of Roncal et al. (2010, 2011). Except for *G. fosteri*, the species of this clade share the character state of apiculate and lobed proximal lips of the flower pits. This is one of the taxonomically most complex groups of the genus, and species delimitation has been handled differently over time (Henderson et al., 1995; Henderson, 2011). In our study, the samples assigned to the widespread species *G. orbignyana* and *G. undata* were not recovered as phylogenetically independent lineages (**Figures 3**, **4**). Rather than clustering according to taxonomic delimitations, specimens of these two species grouped with strong support by geographic location, forming two main clades (each with BS of 100, **Figure 4** and LPP of 1, **Figure 3**): (a) a central American – north Andean clade composed of specimens from Mexico, Costa-Rica, Panama, the Caribbean, and Colombia; and (b) a north Andean – central Andean clade composed of specimens from Ecuador, Peru, and Bolivia. However, in addition to these two subclades, there was also a separate group at the base of the clade, comprising three specimens from Ecuador and Bolivia in the coalescent analysis (with LPP of 1, **Figure 3**) or these three specimens plus *G. trigona* in the concatenated analysis (with BS of 63, **Figure 4**). The samples of *G. lehmannii* subsp. *corrugata* A.J. Hend. and *G. talamancana*, both occurring in Central America, were nested within the central American – north Andean clade with strong support (BS of 100, **Figure 4** and LPP of 1, **Figure 3**). As these two species have resembling morphologies and similar high elevation habitats to *G. undata* and *G. orbignyana*, they may represent locally divergent populations of this broad species complex, in which a novel phenotype has been fixed (e.g., the thick corrugated leaves and the well-developed peduncle in *G. lehmannii* subsp. *corrugata)* although it is not underpinned by strong genetic isolation. Another similar case is *G. trigona,* which, together with *G. fosteri* is recovered as sister to the rest of Clade X in the coalescent analysis (with LPP of 0.8, **Figure 3**). The placement of *G. fosteri* is intriguing because it is morphologically clearly different from the rest of the species in this clade, even though it occurs in the same habitat. Although high-quality DNA was recovered for this sample and the identification of the living specimen from which it was collected was double-checked, we cannot rule out contamination during laboratory work to explain this surprising result.

Clade XI

This clade includes five species (*G. cuneata* H. Wendl. ex Spruce, *G. lanata* A.J. Hend, Borchs & Balslev, *G. laxiflora* Mart., *G. stricta* (Poit.) Kunth, *G. tenuissima* H.E. Moore, **Figure 5**) that Henderson (2011) placed in several distinct clades (*G. cuneata*, *G. lanata*, and *G. stricta* clades). These five species were part of an unresolved clade recovered in Roncal et al. (2012). *Geonoma stricta* and *G. cuneata* are two species complexes whereas the three remaining species of this clade have rather low variability (Henderson, 2011). Henderson (2011) recognized nine subspecies in *G. stricta,* with the most widespread and morphologically variable, subspecies *arundinacea*, further divided into eight morphotypes. Of the three subspecies included in our analysis, subspecies *stricta* and *montana* were nested within *arundinacea*. For the latter, the two morphotypes sampled (*trailii* and *pycnostachys*) were paraphyletic. In *G. cuneata,* nine subspecies were recognized and the most widespread and morphologically variable, subspecies *cuneata* was divided into 13 morphotypes (Henderson, 2011). We did not sample enough individuals of the different subspecies of *G. cuneata* to comment on their recognition using phylogenomics, except that subspecies *guanacastensis* was nested within subspecies *cuneata*. Species of this clade are mostly distributed in lowland and lower montane forests with *G. cuneata* distributed from Nicaragua to

Ecuador, *G. lanata* and *G. tenuissima* occurring mostly on the western Andean slopes, *G. laxiflora* in western Amazonia, and the widespread *G. stricta* overlapping the previous four ranges and reaching central Amazonia and the Guyanas. Although these species are morphologically quite different from each other, which explains why Henderson (2011) recovered them in different clades, they commonly have cane-like stems with yellowish and smooth internodes.

Clade XII

This clade includes a single species, *G. divisa* H.E. Moore (**Figure 5**), which is endemic to northwestern Colombia. Henderson (2011) placed this species in his *G. stricta* clade, alongside *G. longivaginata* H.Wendl. ex Spruce and *G. ferruginea* H. Wendl. ex Spruce, but this relationship was not supported at all in our analyses. Morphologically, *G. divisa* differs from these two species in its tricussately arranged, closely spaced flower pits (Henderson, 2011).

Clade XIII

This clade includes two species (*G. brongniartii* Mart., *G. poeppigiana* Mart.; **Figure 5**) that occur from Colombia to Bolivia. These closely related species are variable and their separation has long been debated (Henderson, 2011). In our phylogenetic tree, these two morphological taxa are mixed together within a single clade rather than forming distinct monophyletic groups (**Figures 3**, **4**). Although we can not rule out sample misidentification, this result could also be explained by alternative hypotheses. Indeed, the geographic distributions of these two species overlap, with *G. poeppigiana* having a smaller range than *G. brongniartii*, and Henderson (2011) reported putative hybrids between them in central Peru. All specimens of *G. poeppigiana* included in our phylogeny come from Peru and so do most specimens of *G. brongniartii*. If the two species indeed hybridize in the Peruvian area where they co-occur, this may explain why they are mixed in our phylogenetic tree. Alternatively, these samples could represent a single widely variable species (see below).

Clade XIV

This clade includes eight species (*G. brenesii* Grayum, *G. epetiolata* H.E. Moore, *G. ferruginea*, *G. hugonis* Grayum & de Nevers, *G. longivaginata*, *G. monospatha* de Nevers, *G. mooreana* de Nevers & Grayum, *G. scoparia* Grayum & de Nevers; **Figure 5**) from Costa Rica and Panama. It corresponds to Roncal et al.'s (2011) central American clade, whereas in Henderson's (2011) tree these species were placed in three different clades (namely the *G. cuneata*, *G. lanata*, and *G. stricta* clades). The three samples of *G. monospatha* from Costa Rica are placed far apart from the two Panamanian samples in our phylogenetic tree (with BS of 100, **Figure 4** and LPP of 1, **Figure 3**). These populations are geographically disjunct, with a gap of several hundred kilometers between them. Furthermore, the Costa Rican specimens have smaller leaves and inflorescences and thrive at higher elevations (mean elevations 1750 m vs. 837 m). All of this suggests that the Costa Rican population may better be treated as a distinct species. Similarly, the two samples of the Panamanian *G. longivaginata* subspecies *copensis* A. J. Hend. are more closely related to the two Panamanian samples of *G. monospatha* (with BS of 100, **Figure 4** and LPP of 1, **Figure 3**) than to the individuals of *G. longivaginata* from Costa Rica, suggesting that this subspecies may, in fact, also better be treated as a distinct species.

Comparison With Other Phylogenetic Reconstructions

In the latest revision of *Geonoma*, Henderson (2011) conducted a maximum parsimony phylogenetic analysis of all the species of the genus based on 30 qualitative morphological characters. We used his species-level taxonomy to name our taxa and tried to apply his clade definition to the phylogenetic tree we obtained. We found that many of his clades were supported by our study, although often with the exclusion or inclusion of a few species. For instance, his *G. cuneata* clade (minus *G. cuneata*) corresponds to our Clade XIV, his *G. macrostachys* clade (minus *G. maxima*) corresponds to our Clade III, his *G. schottiana* clade corresponds to our Clade VII, his *G. undata* clade (plus *G. fosteri*) corresponds to our Clade X, his *G. congesta* clade (plus *G. galeanoae*) to our Clade II, and his *G. interrupta* clade (plus *G. santaderensis*) to our Clade IV. In contrast, Henderson's *G. lanata* and *G. stricta* clades are not supported at all by our analyses, suggesting that these clades were defined by homoplasic morphological characters with limited phylogenetic information. These characters included rachillae surfaces with spiky, fibrous projections or ridges for the *G. stricta* clade, and filiform rachillae with extended narrowed sections between the alternately arranged flower pits for the *G. lanata* clade. Furthermore, the relationships between the clades as recovered by Henderson (2011) are in strong disagreement with the topology obtained from molecular data, indicating that morphology provides little insight on the deep phylogenetic relationships within the genus.

The first molecular phylogeny of *Geonoma* was based on 20 species and two markers (Roncal et al., 2005). It was later extended to three genes and 43 species (Roncal et al., 2010, 2011, 2012). Using an extended sampling of 57 species and 795 gene regions, our study confirmed many of the findings of these studies for the phylogenetic relationships at intermediate levels of divergence. For instance, our Clades I-III, which together are sister to the other 11 clades, correspond to Roncal's (2011) Amazon clade, which was also recovered as sister to the remainder of the genus. Furthermore, the internal arrangements of the species in this group are also largely congruent, with *G. maxima* sister to the remainder of the species in the Amazon clade, and *G. calyptrogynoidea* and *G. congesta* (our Clade II) sister to the remainder of the species (our Clade III), although *G. baculifera* and *G. concinna* were recovered by Roncal et al. (2010) to be more closely related to species in our Clade III than they were to species in Clade II, where we placed them. Likewise, Roncal et al.'s (2011) Brazilian Cerrado + Mata Atlantica, Andes + Central American Mountains, and Central America clades were also recovered in our phylogenetic tree and the relative arrangements of these clades are overall similar between

both studies. In general, previously unresolved phylogenetic relationships were resolved with strong support in our analyses.

Species Delimitation

Henderson (2011) used a statistical approach in which species delimitation was based on the results of a clustering analysis of qualitative morphological characters. The congruence between the results of this method and those of our phylogenetic analysis is striking. Indeed, the majority of species and species complexes for which we included several samples were recovered as monophyletic units in our phylogenetic trees. Furthermore, morphological variation seemed to be correlated to genetic divergence, as indicated by the longer branches of widely variable species in the maximum likelihood tree compared to species with smaller geographic ranges and less variable phenotypes (**Figure 4**). Although further sampling would be necessary in order to include the full range of variability of some widely distributed species, our study nevertheless supports the validity of the characters used by Henderson (2011) to define species boundaries. Conversely, even though not all our samples had identification at the intraspecific level, our results indicate that the intraspecific divisions as subspecies, varieties or morphotypes generally do not match the genetic clusters. These intraspecific taxa were defined by Henderson (2011) usually based on a few, often variable characters such as the degree of division in the inflorescences. Only in *G. pinnatifrons* are the subspecies, which were delimited based on geographical distribution, supported by the tree topology (**Figures 3**, **4**). In other species complexes (e.g., in *G. cuneata*, *G. maxima, G. stricta*, *G. macrostachys*, and *G. pohliana)*, intraspecific taxa are not recovered as monophyletic (**Figures 3**, **4**). Furthermore, while our phylogenetic tree revealed a broad North-South differentiation within the mixed *G. orbignyana* – *G. undata* group, in general there seems to be no geographic clustering of individuals, especially for widespread Amazonian taxa such as *G. macrostachys*, *G. maxima* or *G. stricta*. Various hypotheses have been proposed to explain this kind of chaotic intraspecific phenotypic variation, such as population contraction and expansion during Pleistocene climatic oscillations (Cronk, 1998), rapid dispersal followed by selection (Cronk, 1998), or niche divergence induced by forest heterogeneity (Henderson, 2011). From a genetic perspective, incomplete lineage sorting or hybridization are commonly invoked to explain the widespread occurrence of plant species complexes similar to those found in *Geonoma* (Bacon et al., 2012b; Pinheiro et al., 2018). Although our results pointed to high level of gene tree incongruence in the species complexes of *Geonoma* (**Figure 3**), likely due to incomplete lineage sorting, it is beyond the scope of our study to test for any of these underlying mechanisms. Further phylogeographic or population genetic studies are needed to understand the origin of the discrepancy between morphological and genetic data in this group.

From a systematic point of view, the remaining issue to be addressed is the taxonomic status of the several non-monophyletic species that were identified by our analysis. First, there are two cases where two species were recovered mixed within a single clade (*G. brongniartii* with *G. poeppigiana* and *G. orbignyana* with *G. undata*). Second, there are several instances of geographically restricted species (e.g., *G. lehmanii, G. poiteauana, G. talamancana*, and *G. trigona*) which were found to be nested within more widely distributed species, making the latter paraphyletic. For taxonomic classification, there are two fundamentally different approaches to deal with such situations. On one hand, under a lineage species concept, which requires species monophyly, the phylogenetically intermixed "species" of *Geonoma* would be considered to represent single variable species as was done in other similar cases in plants (Bennett et al., 2008; Barbosa et al., 2012). Applying this approach would entail a reduction in the number of recognized species within *Geonoma*. On the other hand, some authors have stressed that at the species level paraphyly is not an issue because it is considered to be the natural output of, for example, peripatric speciation or speciation via polyploidy which are common phenomena in plants (Rieseberg and Brouillet, 1994; Crisp and Chandler, 1996). Likewise, the pattern of small-range species nested within large-range paraphyletic species has been suggested to be common in rainforest trees with widespread distribution for which coalescence times are long due to large population size and extensive gene flow (Pennington and Lavin, 2016). Non-monophyletic species can also arise from hybridization, which is widespread in plants (Whitney et al., 2010). Under a genic species concept, species cohesion may be determined by a small number of genes, thus allowing gene flow between species without calling species identity into question (Wu, 2001; Lexer and Widmer, 2008). Therefore, adopting a species concept which places emphasis on the phenotype would result in treating morphologically divergent entities as separate species even if they do no represent evolutionary independent lineages (Freudenstein et al., 2016). From this point of view, the number of species in *Geonoma* would remain similar to that proposed by Henderson (2011). In the end, how the phylogenetic information presented here is translated into a taxonomic classification is to a certain degree a matter of personal preference, with different researchers favoring different aspects. Thus, some may emphasize morphological or genetic similarities while others would focus on differences, some place more importance on the ability to diagnose taxa while others prioritize evolutionary independence, and so on. We refrain from proposing taxonomic decisions based on our results, since this would require a full assessment of genetic, morphological, and ecological evidence.

CONCLUSION

By employing a large novel set of molecular markers, we were able to clarify both deep and shallow phylogenetic relationships within the tribe Geonomateae including for *Geonoma,* one of the largest and taxonomically most challenging Neotropical palm genera. The remaining poorly supported phylogenetic relationships do not reflect a lack of informative genetic data but are rather caused by a high level of gene tree incongruence, as shown by the coalescent analysis. Our phylogenetic analyses revealed two cryptic species of *Geonoma* in Central America, which will have to be described in further taxonomic work. The intraspecific sampling confirmed in most cases the validity of the taxonomic delimitation of species proposed by Henderson (2011),

even for those with extensive phenotypic variability such as *G. cuneata*, *G. interrupta*, *G. maxima*, *G. pinnatifrons*, *G. macrostachys*, or *G. stricta*. However, we also pointed to several cases where the morphological delimitations do not reflect the genetic clusters, such as the internal delimitations of widely variable species complexes, the clustering of rare endemic species within broader species complexes, or the mixing of two species complexes. These groups that do not show clear genetic boundaries between morphologically recognized taxa remain the main challenge in the systematics of *Geonoma*. Ultimately, the number of species recognized in *Geonoma* depends on the species concept one endorses.

Studies at the population level are needed to understand whether the decoupling between morphological and genetic variation in the species complexes is the result of ongoing speciation with gene flow or from secondary contact and hybridization between previously diverged taxa. Although the impossibility of summarizing morphological variation of these groups into a coherent classification scheme may seem frustrating from a taxonomic point of view, we argue that it represents a unique opportunity to better understand the build-up of Neotropical plant diversity. Indeed, species complexes are common in plants and are gaining attention as model groups to study the underlying factors of plant speciation (Pinheiro et al., 2018). In this context, the set of baits recently developed by de La Harpe et al. (2019) will therefore be a useful tool to carry out specific population levels studies.

With this in mind, we provided the baits for a selection including 20% of the most informative genes from the kit developed by de La Harpe et al. (2019) by assessing their phylogenetic informativeness across three Arecaceae subfamilies. We predict that these baits should work over a wide evolutionary timescale in the Arecaceae and will therefore benefit the whole field of palm phylogenomics. Indeed, the smaller size of this kit will make it accordingly more affordable and will reduce the computation time of post-sequencing bioinformatic analyses while maximizing the phylogenetic informativeness at deep and shallow scales across the Arecaceae family. Hence, we believe that it has the potential to be an essential tool in the search toward a complete species-level phylogeny of the Arecaceae family.

DATA AVAILABILITY

Targeted sequence reads generated as part of this manuscript are available in NCBI (BioProject PRJNA541164). The list of the 795 selected genes and their corresponding bait sequences in fasta format as well as all gene trees and species trees were deposited in Zenodo (deposit number 2594808).

AUTHOR CONTRIBUTIONS

NS, MK, and CL designed the study. MPa led the sequencing experiment and performed the post-sequencing bioinformatics analyses. MdLH, OL, TM-M, and MPa did the labwork. OL and DK performed the phylogenetic analyses. CB, HB, FB, AC, TC, MPe, JuR, MS, FS, CL, MK, and NS were part of the Geonoma Consortium set up to perform this study. OL led the writing with significant contributions from all co-authors, in particular MK, IO, and NS. All co-authors commented and agreed on the last version of the manuscript.

FUNDING

NS, MK, and CL received funding from the Swiss National Science Foundation (CRSII3-147630), NS from the University of Lausanne, JoR from a Banting postdoctoral fellowship (151042) at University of British Columbia, MPe from the Swiss National Science Foundation (Grant No. 31003A_175655/1), TM-M from the CNPq-SWE (205660/2014-2), MS from the Colciencias (Contract No. 173-2016), and HB from the Danish Council for Independent Research – Natural Sciences (4181-00158) and the European Community (FP7 212631).

ACKNOWLEDGMENTS

This study would not have been possible without Andrew Henderson's help. We are deeply grateful to him for identifying many of the samples and for his invaluable comments on an earlier version of the manuscript. We thank Natalia Arcila, Tatiana Boza, Lilian Costa Procopio, Camilo Flórez, María Fernanda González, Rosa Isela Meneses, Vanessa Rojas, Alain Rousteau, Lázaro Santa Cruz, Adrian Tejedor, Johanna Toivonen, and Erickson Urquiaga for their assistance in the field. Gloria Galeano[5] and Jean Christophe Pintaud[5] for their assistance with fieldwork planning and the identification of some specimens. AAU and COL for providing silica-dried leaves of many samples; G and COL for herbarium fragments; and Rodrigo Bernal and Juan Carlos Copete for providing photos of some vouchers. We also thank the Vital-IT facilities of the Swiss Institute of Bioinformatics for the use of their HPC infrastructure.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.00864/full#supplementary-material

[5] Deceased.

REFERENCES

Abadi, S., Azouri, D., Pupko, T., and Mayrose, I. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nat. Commun.* 10:934. doi: 10.1038/s41467-019-08822-w

Al-Mssallem, I. S., Hu, S., Zhang, X., Lin, Q., Liu, W., Tan, J., et al. (2013). Genome sequence of the date palm *Phoenix dactylifera* L. *Nat. Commun.* 4:2274. doi: 10.1038/ncomms3274

Bacon, C. D., Baker, W. J., and Simmons, M. P. (2012a). Miocene dispersal drives island radiations in the palm tribe Trachycarpeae

(Arecaceae). *Syst. Biol.* 61, 426–442. doi: 10.1093/sysbio/syr123

Bacon, C. D., McKenna, M. J., Simmons, M. P., and Wagner, W. L. (2012b). Evaluating multiple criteria for species delimitation: an empirical example using Hawaiian palms (Arecaceae: *Pritchardia*). *BMC Evol. Biol.* 12:23. doi: 10.1186/1471-2148-12-23

Bacon, C. D., Look, S. L., Gutiérrez-Pinto, N., Antonelli, A., Tan, H. T., Kumar, P. P., et al. (2016a). Species limits, geographical distribution and genetic diversity in *Johannesteijsmannia* (Arecaceae). *Bot. J. Linn. Soc.* 182, 318–347. doi: 10.1111/boj.12470

Bacon, C. D., Velásquez-Puentes, F., Flórez-Rodríguez, A., Balslev, H., Galeano, G., Bernal, R., et al. (2016b). Phylogenetics of Iriarteeae (Arecaceae), cross-Andean disjunctions and convergence of clustered infructescence morphology in *Wettinia*. *Bot. J. Linn. Soc.* 182, 272–286. doi: 10.1111/boj.12421

Bacon, C. D., Moraes, M., Jaramillo, C., and Antonelli, A. (2017). Endemic palm species shed light on habitat shifts and the assembly of the Cerrado and Restinga floras. *Mol. Phylogenet. Evol.* 110, 127–133. doi: 10.1016/j.ympev.2017.03.013

Baker, W. J., Asmussen, C. B., Barrow, S. C., Dransfield, J., and Hedderson, T. A. (1999). A phylogenetic study of the palm family (Palmae) based on chloroplast DNA sequences from the trnL - trnF region. *Plant Syst. Evol.* 219, 111–126. doi: 10.1007/BF01090303

Baker, W. J., and Couvreur, T. L. P. (2013). Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. II. Diversification history and origin of regional assemblages. *J. Biogeogr.* 40, 286–298. doi: 10.1111/j.1365-2699.2012.02794.x

Baker, W. J., and Dransfield, J. (2016). Beyond *Genera Palmarum*: progress and prospects in palm systematics. *Bot. J. Linn. Soc.* 182, 207–233. doi: 10.1111/boj.12401

Baker, W. J., Norup, M. V., Clarkson, J. J., Couvreur, T. L. P., Dowe, J. L., Lewis, C. E., et al. (2011). Phylogenetic relationships among arecoid palms (Arecaceae: Arecoideae). *Ann. Bot.* 108, 1417–1432. doi: 10.1093/aob/mcr020

Baker, W. J., Savolainen, V., Asmussen-Lange, C. B., Chase, M. W., Dransfield, J., Forest, F., et al. (2009). Complete generic-level phylogenetic analyses of palms (arecaceae) with comparisons of supertree and supermatrix approaches. *Syst. Biol.* 58, 240–256. doi: 10.1093/sysbio/syp021

Balslev, H., Copete, J., Pedersen, D., Bernal, R., Galeano, G., Duque, Á, et al. (2017). "Palm diversity and abundance in the Colombian Amazon," in *Forest structure, function and dynamics in Western Amazonia*, ed. R. W. Myster (London: John Wiley & Sons Ltd.).

Balslev, H., Laumark, P., Pedersen, D., and Grández, C. (2016). Tropical rainforest palm communities in Madre de Dios in Amazonian Peru. *Rev. Peru. Biol.* 23, 3–12.

Barbosa, A. R., Fiorini, C. F., Silva-Pereira, V., Mello-Silva, R., and Borba, E. L. (2012). Geographical genetic structuring and phenotypic variation in the *Vellozia hirsuta* (Velloziaceae) ochlospecies complex. *Am. J. Bot.* 99, 1477–1488. doi: 10.3732/ajb.1200070

Barrett, C., Sinn, B., King, L., Medina, J., Bacon, C., Lahmeyer, S., et al. (2018). Phylogenomics, biogeography, and evolution in the American palm genus *Brahea*. *bioRxiv* 467779. doi: 10.1101/467779

Barrett, C. F., Bacon, C. D., Antonelli, A., and Hofmann, T. (2016). An introduction to plant phylogenomics with a focus on palms. *Bot. J. Linn. Soc.* 182, 234–255. doi: 10.1111/boj.12399

Bennett, J. R., Wood, J. R. I., and Scotland, R. W. (2008). Uncorrelated variation in widespread species: species delimitation in *Strobilanthes echinata* Nees (Acanthaceae). *Bot. J. Linn. Soc.* 156, 131–141. doi: 10.1111/j.1095-8339.2007.00756.x

Blom, M. P. K., Bragg, J. G., Potter, S., and Moritz, C. (2017). Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. *Syst. Biol.* 66, 352–366. doi: 10.1093/sysbio/syw089

Borchsenius, F. (2002). Staggered flowering in four sympatric varieties of *Geonoma cuneata* (Palmae). *Biotropica* 34, 603–606. doi: 10.1111/j.1744-7429.2002.tb00580.x

Borchsenius, F., Lozada, T., and Knudsen, J. T. (2016). Reproductive isolation of sympatric forms of the understorey palm *Geonoma macrostachys* in western Amazonia. *Bot. J. Linn. Soc.* 182, 398–410. doi: 10.1111/boj.12428

Chazdon, R. L. (1991). Plant size and Form in the understory palm genus *Geonoma*: are species variations on a theme? *Am. J. Bot.* 78, 680–694. doi: 10.1002/j.1537-2197.1991.tb12592.x

Chazdon, R. L. (1992). Patterns of growth and reproduction of *Geonoma congesta*, a clustered understory palm. *Biotropica* 24, 43–51.

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

Comer, J. R., Zomlefer, W. B., Barrett, C. F., Davis, J. I., Stevenson, D. W., Heyduk, K., et al. (2015). Resolving relationships within the palm subfamily Arecoideae (Arecaceae) using plastid sequences derived from next-generation sequencing. *Am. J. Bot.* 102, 888–899. doi: 10.3732/ajb.1500057

Comer, J. R., Zomlefer, W. B., Barrett, C. F., Stevenson, D. W., Heyduk, K., and Leebens-Mack, J. H. (2016). Nuclear phylogenomics of the palm subfamily Arecoideae (Arecaceae). *Mol. Phylogenet. Evol.* 97, 32–42. doi: 10.1016/j.ympev.2015.12.015

Couvreur, T. L. P., Forest, F., and Baker, W. J. (2011). Origin and global diversification patterns of tropical rain forests: inferences from a complete genus-level phylogeny of palms. *BMC Biol.* 9:44. doi: 10.1186/1741-7007-9-44

Crisp, M. D., and Chandler, G. T. (1996). Paraphyletic species. *Telopea* 6, 813–844. doi: 10.7751/telopea19963037

Cronk, Q. (1998). "The ochlospecies concept," in *Chorology, Taxonomy and Ecology of the Floras of Africa and Madagascar*, eds C. R. Huxley, J. M. Lock, and D. F. Cutler (Kew: Royal Botanic Gardens), 155–170.

Cuenca, A., Asmussen-Lange, C. B., and Borchsenius, F. (2008). A dated phylogeny of the palm tribe Chamaedoreeae supports Eocene dispersal between Africa, North and South America. *Mol. Phylogenet. Evol.* 46, 760–775. doi: 10.1016/j.ympev.2007.10.010

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

de La Harpe, M., Hess, J., Loiseau, O., Salamin, N., Lexer, C., and Paris, M. (2019). A dedicated target capture approach reveals variable genetic markers across micro-and macro-evolutionary time scales in palms. *Mol. Ecol. Resour.* 19, 221–234. doi: 10.1111/1755-0998.12945

Dransfield, J., Uhl, N. W., Asmussen, C. B., Baker, W. J., Harley, M. M., and Lewis, C. E. (2008). *Genera Palmarum: The Evolution and Classifi Cation of Palms*, 2nd Edn. Kew: Royal Botanical Gardens.

Faircloth, B. C., Chang, J., and Alfaro, M. E. (2012). TAPIR enables high-throughput estimation and comparison of phylogenetic informativeness using locus-specific substitution models. *arXiv preprint* arXiv:1202.1215.

Freudenstein, J. V., Broe, M. B., Folk, R. A., and Sinn, B. T. (2016). Biodiversity and the species concept—lineages are not enough. *Syst. Biol.* 66, 644–656. doi: 10.1093/sysbio/syw098

Gruca, M., Blach-Overgaard, A., and Balslev, H. (2015). African palm ethnomedicine. *J. Ethnopharmacol.* 165, 227–237. doi: 10.1016/j.jep.2015.02.050

Henderson, A. (2002). *Evolution and Ecology of Palms*. New York, NY: New York Botanical Garden Press.

Henderson, A. (2005). A multivariate study of *Calyptrogyne* (Palmae). *Syst. Bot.* 30, 60–83. doi: 10.1600/0363644053661913

Henderson, A. (2011). A revision of *Geonoma* (Arecaceae). *Phytotaxa* 17, 1–271.

Henderson, A. (2012). A revision of *Pholidostachys* (Arecaceae). *Phytotaxa* 43, 1–48.

Henderson, A., Galeano, G., and Bernal, R. (1995). *A Field Guide to the Palms of the Americas*. Princeton, NJ: Princeton University Press.

Henderson, A., and Martins, R. (2002). Classification of specimens in the *Geonoma stricta* (Palmae) complex: the problem of leaf size and shape. *Brittonia* 54, 202–212. doi: 10.1663/0007-196x(2002)054%5B0202:cositg%5D2.0.co;2

Henderson, A., and Villalba, I. (2013). A revision of *Welfia* (Arecaceae). *Phytotaxa* 119, 33–44. doi: 10.11646/phytotaxa.119.1.3

Heyduk, K., Trapnell, D. W., Barrett, C. F., and Leebens-mack, J. I. M. (2015). Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biol. J. Linn. Soc.* 117, 106–120. doi: 10.1111/bij.12551

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

Hoff, M., Orf, S., Riehm, B., Darriba, D., and Stamatakis, A. (2016). Does the choice of nucleotide substitution models matter topologically? *BMC Bioinformatics* 17:143. doi: 10.1186/s12859-016-0985-x

Kircher, M., Sawyer, S., and Meyer, M. (2011). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40:e3. doi: 10.1093/nar/gkr771

Knudsen, J. T., Andersson, S., Bergman, P., Bergman, P., and Knudsen, J. T. (1998). Floral scent attraction in *Geonoma macrostachys*, an understorey palm of the Amazonian rain forest. *Oikos* 85, 409–418.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lexer, C., and Widmer, A. (2008). The genic view of plant speciation: recent progress and emerging questions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 3023–3036. doi: 10.1098/rstb.2008.0078

Ley-lopez, J. M., and Avalos, G. (2017). Propagation of the palm flora in a lowland tropical rainforest in Costa Rica: fruit collection and germination patterns. *Trop. Conserv. Sci.* 10, 1–12. doi: 10.1177/1940082917740703

Liu, L., Xi, Z., Wu, S., Davis, C. C., and Edwards, S. V. (2015). Estimating phylogenetic trees from genome-scale data. *Ann. N. Y. Acad. Sci.* 1360, 36–53. doi: 10.1111/nyas.12747

Liu, L., Yu, L., Kubatko, L., Pearl, D. K., and Edwards, S. V. (2009). Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.* 53, 320–328. doi: 10.1016/j.ympev.2009.05.033

Macia, M. J., Armesilla, P. J., Cámara-Leret, R., Paniagua-Zambrana, N., Villalba, S., Balslev, H., et al. (2011). Palm uses in northwestern South America: a quantitative review. *Bot. Rev.* 77, 462–570. doi: 10.1007/s12229-011-9086-8

Mandel, J. R., Barker, M. S., Bayer, R. J., Dikow, R. B., Gao, T., 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

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meerow, A. W., Noblick, L., Salas-Leiva, D. E., Sanchez, V., Francisco-Ortega, J., Jestrow, B., et al. (2015). Phylogeny and historical biogeography of the cocosoid palms (Arecaceae, Arecoideae, Cocoseae) inferred from sequences of six WRKY gene family loci. *Cladistics* 31, 509–534. doi: 10.1111/cla.12100

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

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

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

Mitchell, N., Lewis, P. O., Moriarty Lemmon, E., Lemmon, A. R., and Holsinger, K. E. (2017). Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. *Am. J. Bot.* 104, 102–115. doi: 10.3732/ajb.1600227

Muscarella, R., Bacon, C. D., Faurby, S., Antonelli, A., Kristiansen, S. M., Svenning, J.-C., et al. (2018). Soil fertility and flood regime are correlated with phylogenetic structure of Amazonian palm communities. *Ann. Bot.* 123, 641–655. doi: 10.1093/aob/mcy196

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

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

Pennington, R. T., and Lavin, M. (2016). The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *New Phytol.* 210, 25–37. doi: 10.1111/nph.13724

Pinheiro, F., Dantas-queiroz, M. V., and Palma-silva, C. (2018). Plant species complexes as models to understand speciation and evolution: a review of South American studies. *Crit. Rev. Plant Sci.* 37, 54–80. doi: 10.1080/07352689.2018.1471565

Pizo, M. A., and Almeida-Neto, M. (2009). Determinants of fruit removal in *Geonoma pauciflora*, an understory palm of neotropical forests. *Ecol. Res.* 24, 1179–1186. doi: 10.1007/s11284-009-0599-0

Pond, S. L. K., and Muse, S. V. (2005). "HyPhy: hypothesis testing using phylogenies," in *Statistical Methods in Molecular Evolution*, ed. R. Nielsen (New York, NY: Springer New York), 125–181. doi: 10.1007/0-387-27733-1_6

Rambaut, A. (2012). *FigTree v1. 4.* Available at: http://tree.bio.ed.ac.uk/software/figtree/ (accessed 28 April 2014).

Rieseberg, L. H., and Brouillet, L. (1994). Are many plant species paraphyletic? *Taxon* 43, 21–32. doi: 10.2307/1223457

Rodriguez-Buritica, S., Orjuela, M. A., and Galeano, G. (2005). Demography and life history of *Geonoma orbignyana*: an understory palm used as foliage in Colombia. *For. Ecol. Manag.* 211, 329–340. doi: 10.1016/j.foreco.2005.02.052

Roncal, J. (2006). Habitat differentiation of sympatric *Geonoma macrostachys* (Arecaceae) varieties in Peruvian lowland forests. *J. Trop. Ecol.* 22, 483–486. doi: 10.1017/s0266467406003270

Roncal, J., Blach-Overgaard, A., Borchsenius, F., Balslev, H., and Svenning, J. C. (2011). A dated phylogeny complements macroecological analysis to explain the diversity patterns in *Geonoma* (Arecaceae). *Biotropica* 43, 324–334. doi: 10.1111/j.1744-7429.2010.00696.x

Roncal, J., Borchsenius, F., Asmussen-Lange, C. B., and Balslev, H. (2010). "Divergence times in the tribe Geonomateae (Arecaceae) coincide with tertiary geological events," in *Diversity, Phylogeny and Evolution of the Monocotyledons*, eds O. Seberg, G. Petersen, A. S. Barfod, and J. I. Davis (Aarhus: Aarhus University Press), 245–265.

Roncal, J., Francisco-Ortega, J., Asmussen, C. B., and Lewis, C. E. (2005). Molecular phylogenetics of tribe Geonomeae (Arecaceae) using nuclear DNA sequences of Phosphoribulokinase and RNA Polymerase II. *Syst. Bot.* 30, 275–283. doi: 10.1600/0363644054223620

Roncal, J., Francisco-Ortega, J., and Lewis, C. E. (2007). An evaluation of the taxonomic distinctness of two *Geonoma macrostachys* (Arecaceae) varieties based on intersimple sequence repeat (ISSR) variation. *Bot. J. Linn. Soc.* 153, 381–392. doi: 10.1111/j.1095-8339.2007.00619.x

Roncal, J., Henderson, A., Borchsenius, F., Cardoso, S. R. S., and Balslev, H. (2012). Can phylogenetic signal, character displacement, or random phenotypic drift explain the morphological variation in the genus *Geonoma* (Arecaceae)? *Biol. J. Linn. Soc.* 106, 528–539. doi: 10.1111/j.1095-8312.2012.01879.x

Roncal, J., Zona, S., and Lewis, C. E. (2008). Molecular phylogenetic studies of Caribbean palms (Arecaceae) and their relationships to biogeography and conservation. *Bot. Rev.* 74, 78–102. doi: 10.1007/s12229-008-9005-9

Sampaio, M. B., and Scariot, A. (2008). Growth and reproduction of the understory palm *Geonoma schottiana* Mart. in the gallery forest in Central Brazil. *Rev. Bras. Bot.* 31, 433–442. doi: 10.1590/S0100-84042008000300007

Sanín, M. J., Kissling, W. D., Bacon, C. D., Borchsenius, F., Galeano, G., Svenning, J.-C., et al. (2016). The neogene rise of the tropical Andes facilitated diversification of wax palms (*Ceroxylon*: Arecaceae) through geographical colonization and climatic niche separation. *Bot. J. Linn. Soc.* 182, 303–317. doi: 10.1111/boj.12419

Sass, C., Iles, W. J. D., Barrett, C. F., Smith, S. Y., and Specht, C. D. (2016). Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* 4:e1584. doi: 10.7717/peerj.1584

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

Singh, R., Ong-Abdullah, M., Low, E.-T. L., Manaf, M. A. A., Rosli, R., Nookiah, R., et al. (2013). Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature* 500, 335–339. doi: 10.1038/nature12309

Smeds, L., and Künstner, A. (2011). ConDeTri-a content dependent read trimmer for Illumina data. *PLoS One* 6:e26314. doi: 10.1371/journal.pone.0026314

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

Stauffer, F. W. (1997). Estudio morfologico y taxonomico de *Geonoma spinescens* H. Wendl. ex Burret (Arecaceae) y descripcion de una nueva variedad. *Acta Bot. Venezuelica* 20, 1–10.

Stauffer, F. W., Asmussen, C. B., Henderson, A., and Endress, P. K. (2003). A revision of *Asterogyne* (Arecaceae: Arecoideae: Geonomeae). *Brittonia* 55, 326–356. doi: 10.1663/0007-196x(2003)055%5B0326:aroaaa%5D2.0.co;2

Townsend, J. P. (2007). Profiling phylogenetic informativeness. *Syst. Biol.* 56, 222–231. doi: 10.1080/10635150701311362

Uhl, N. W., and Dransfield, J. (1987). *Genera Palmarum. A Classification of Palms Based on the Work of Harold E. Moore, Jr*. Lawrence, KS: Allen Press.

Uthaipaisanwong, P., Chanprasert, J., Shearman, J. R., Sangsrakru, D., Yoocha, T., Jomchai, N., et al. (2012). Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq.). *Gene* 500, 172–180. doi: 10.1016/j.gene.2012.03.061

Vormisto, J., Svenning, J., Hall, P., and Balslev, H. (2004). Diversity and dominance in palm (Arecaceae) communities in terra firme forests in the western Amazon basin. *J. Ecol.* 92, 577–588. doi: 10.1111/j.0022-0477.2004.00904.x

Wessels Boer, J. G. (1968). The geonomoid palms. *Meded. Van Het Bot. Mus. En Herb. Van Rijksuniv. Te Utrecht* 282, 1–202. doi: 10.3732/ajb.89.11.1772

Whitney, K. D., Ahern, J. R., Campbell, L. G., Albert, L. P., and King, M. S. (2010). Patterns of hybridization in plants. *Perspect. Plant Ecol. Evol. Syst.* 12, 175–182.

Wilson, M. A., Gaut, B., and Clegg, M. T. (1990). Chloroplast DNA evolves slowly in the palm family (Arecaceae). *Mol. Biol. Evol.* 7, 303–314.

Wu, C. (2001). The genic view of the process of speciation. *J. Evol. Biol.* 14, 851–865.

Yang, M., Zhang, X., Liu, G., Yin, Y., Chen, K., Yun, Q., et al. (2010). The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). *PLoS One* 5:e12762. doi: 10.1371/journal.pone.0012762

Zona, S. (1995). A revision of *Calyptronoma*. *Principes* 39, 140–151.

Conflict of Interest Statement: 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: 17 September 2019
doi: 10.3389/fpls.2019.01074

Phylogeny of Hawaiian *Melicope* (Rutaceae): RAD-seq Resolves Species Relationships and Reveals Ancient Introgression

Claudia Paetzold[1], Kenneth R. Wood[2], Deren A. R. Eaton[3,4], Warren L. Wagner[5] and Marc S. Appelhans[1,5]*

[1] Department of Systematics, Biodiversity and Evolution of Plants (with Herbarium), University of Göttingen, Goettingen, Germany, [2] National Tropical Botanical Garden, Kalaheo, HI, United States, [3] Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, NY, USA, [4] Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY, United States, [5] Department of Botany, Smithsonian Institution, Washington, DC, United States

Hawaiian *Melicope* are one of the major adaptive radiations of the Hawaiian Islands comprising 54 endemic species. The lineage is monophyletic with an estimated crown age predating the rise of the current high islands. Phylogenetic inference based on Sanger sequencing has not been sufficient to resolve species or deeper level relationships. Here, we apply restriction site-associated DNA sequencing (RAD-seq) to the lineage to infer phylogenetic relationships. We employ Quartet Sampling to assess information content and statistical support, and to quantify discordance as well as partitioned ABBA-BABA tests to uncover evidence of introgression. Our new results drastically improved resolution of relationships within Hawaiian *Melicope*. The lineage is divided into five fully supported main clades, two of which correspond to morphologically circumscribed infrageneric groups. We provide evidence for both ancestral and current hybridization events. We confirm the necessity for a taxonomic revision of the *Melicope* section *Pelea*, as well as a re-evaluation of several species complexes by combining genomic and morphological data.

Keywords: adaptive radiation, *D*-statistics, Hawaiian flora, introgression, Marquesas Islands, Quartet Sampling, RAD-seq, Rutaceae

OPEN ACCESS

Edited by:
Gonzalo Nieto Feliner, Real Jardín Botánico (RJB), Spain

Reviewed by:
Ya Yang, University of Minnesota Twin Cities, United States
Mario Fernández-Mazuecos, Real Jardín Botánico (RJB), Spain

****Correspondence:***
Claudia Paetzold
claudia.paetzold@uni-goettingen.de

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

Received: *24 April 2019*
Accepted: *07 August 2019*
Published: *17 September 2019*

Citation:
Paetzold C, Wood KR, Eaton DAR, Wagner WL and Appelhans MS (2019) Phylogeny of Hawaiian Melicope (Rutaceae): RAD-seq Resolves Species Relationships and Reveals Ancient Introgression. Front. Plant Sci. 10:1074. doi: 10.3389/fpls.2019.01074

INTRODUCTION

Oceanic islands have long been a focal point of evolutionary studies, as they represent a microcosm for examining the process of speciation. This microcosm is shaped by a combination of factors: (1) islands are geographically small and discrete units, sometimes far removed from continental landmasses; (2) colonizations or secondary arrivals are relatively infrequent, and thus, gene flow between the source areas and island systems is restricted; and (3) islands often have dynamic geological histories that give rise to extensively varying landscapes with numerous ecological niches (Emerson, 2002; Price and Wagner, 2018). These factors can often lead to high levels of endemism, which is often the result of adaptive radiation of a limited number of colonizers (Price and Wagner, 2004; Losos and Ricklefs, 2009; Keeley and Funk, 2011). Synthesizing the unique aspects of island evolution and extrapolating results to larger scales may allow us to better uncover general patterns and processes in evolution. Such phenomena include identifying factors affecting

successful colonization and adaptive radiation (Carlquist, 1967; Carlquist, 1974; Paetzold et al., 2018), morphological or ecological shifts (e.g., "insular woodiness"; Carlquist, 1974; Lens et al., 2013), the spatiotemporal origins of lineages (Appelhans et al., 2018a), reconstructing colonization events (Harbaugh et al., 2009), and studying co-evolution (Roderick, 1997). These insights may result in further questions regarding taxonomy, species richness, medicinal or technical applications, and conservation (e.g., Francisco-Ortega et al., 2000).

Adaptive radiations on islands are of special interest for connecting changes in morphology and ecology through time (Givnish, 1998) but require well-resolved phylogenies to do so. In the Hawaiian Islands, phylogenetic studies based on morphology and taxonomy have sometimes overestimated the number of colonization events, because high levels of morphological diversity led researchers to overestimate lineage diversity and the number of colonization events (Price and Wagner, 2018). In contrast, molecular phylogenetic studies have revealed that many enigmatic Hawaiian plant radiations colonized the islands only once followed by adaptive radiation: the Hawaiian lobeliads (Campanulaceae; Givnish et al., 2009), *Psychotria* (Rubiaceae; Nepokroeff et al., 2003), *Silene* (Caryophyllaceae; Eggens et al., 2007), *Touchardia/Urera* (Urticaceae; Wu et al., 2013), and *Melicope* (Harbaugh et al., 2009; Appelhans et al., 2014a). Polyploidization and hybridization events were also discovered to predate colonization and radiation in several island lineages, including the Hawaiian silverswords (Asteraceae; Baldwin and Sanderson, 1998; Barrier et al., 1999) and mints (Lamiaceae; Roy et al., 2015) along with the Pan-Pacific sandalwoods (Santalaceae; Harbaugh, 2008), suggesting evolutionary success in young hybrid or polyploid colonists (Carr, 1998; Paetzold et al., 2018).

Time-scaled phylogenies have revealed that most Hawaiian radiations are ≤5 Myr old, which corresponds to the age of the oldest current main islands, Kaua'i and Ni'ihau. This suggests a bottleneck for dispersal from older (and now largely submerged) leeward islands to the current main islands. However, there are several known exceptions of lineages older than 5 Myr, including *Drosophila*, damselflies, lobeliads, *Zanthoxylum* (Rutaceae), as well as *Melicope* (Price and Clague, 2002; Keeley and Funk, 2011; Appelhans et al., 2018a; Appelhans et al., 2018b). Most phylogenetic studies of Hawaiian flora, however, have relied on few sequenced loci and have thus lacked sufficient power to resolve recent rapid radiations where hybridization, incomplete lineage sorting (ILS), and polyploidy may be common. Newer genomic tools are likely to provide more accurate estimates that may transform our understanding of island radiations.

The genus *Melicope* comprises about 235 species of shrubs and trees distributed throughout SE Asia and Australasia, extending to the Mascarene Islands and Madagascar in the West and most of the Pacific Archipelagos in the East (Hartley, 2001). There are 54 species of *Melicope* endemic to the Hawaiian Islands (Appelhans et al., 2017; Wood et al., 2017), 41 of which are single island endemics (Stone et al., 1999). Hawaiian *Melicope* were initially placed in the genus *Pelea* together with species from the Marquesas Islands (Stone, 1969; Stone et al., 1999) but later incorporated into *Melicope*, forming the majority of the section *Pelea* (Hartley, 2001). Hawaiian *Pelea* was divided into four sections based mainly on the grade of carpel connation: *Apocarpa*, *Cubicarpa*, *Megacarpa*, and *Pelea*. Since the incorporation of the genus *Pelea* into *Melicope*, these sections have not been formally recognized within the larger infrageneric taxonomy for *Melicope* as recognized by Hartley (2001) but are still being used informally as species groups (Appelhans et al., 2014a), and we refer to them as Stone's sectional species groups (Stone's sections) from here on. The most current and comprehensive taxonomic treatment of Hawaiian *Melicope* was considered "provisional" by the authors (Stone et al., 1999), as species boundaries are difficult to define in some cases. Examples include three described species complexes, where the incorporated species vary from each other primarily in the degree of fruit pubescence; the *Melicope elliptica* complex based mainly in O'ahu (six species), the Hawaiian-based *Melicope volcanica* complex (four species), and the Kaua'i-based *Melicope kavaiensis* complex (five species) (Stone et al., 1999).

In contrast to other successful island radiations, the colonization of the Hawaiian Archipelago in *Melicope* was not preceded by a recent polyploidization event. In general, the genus *Melicope* shows a uniform chromosome number (Paetzold et al., 2018). To date, phylogenetic relationships in Hawaiian *Melicope* have been investigated in four molecular studies (Harbaugh et al., 2009; Appelhans et al., 2014a; Appelhans et al., 2014b; Appelhans et al., 2018a), with a combination of up to six nuclear and plastid genomic regions amplified using polymerase chain reaction. Hawaiian *Melicope* was shown to be derived from a single colonization event (Harbaugh et al., 2009). The origin of the lineage was dated to the Mid or Late Miocene (Appelhans et al., 2018a), predating the age of Kaua'i and Ni'ihau (Price and Clague, 2002). In addition, the Hawaiian endemic genus *Platydesma* is nested within *Melicope* as a monophyletic sister group to the Hawaiian species and has since been reduced (Appelhans et al., 2017). Statistically supported incongruences between individual genomic regions were not observed, yet the resolution of relationships within and among the clades was in general medium to poor (Harbaugh et al., 2009; Appelhans et al., 2014a; Appelhans et al., 2014b; Appelhans et al., 2018a). However, two independent Hawaiian origins of the Marquesan *Melicope* radiation, which encompasses seven species, were inferred (Appelhans et al., 2014a; Appelhans et al., 2014b; Appelhans et al., 2017).

Restriction site-associated DNA sequencing (RAD-seq; Miller et al., 2007; Baird et al., 2008) is among the most frequently used reduced representation methods employed in plant systematics. To date, most phylogenetic RAD-seq studies have focused mostly on populations or closely related species (Ree and Hipp, 2015; Díaz-Arce et al., 2016; Hodel et al., 2017). However, a simulated RAD investigation in *Drosophila* revealed the method to be potentially applicable in groups aged up to 60 Myr (Rubin et al., 2012). Since then, application to deeper species-level relationships has increased (e.g., Eaton and Ree, 2013; Hipp et al., 2014; Eaton et al., 2017), facilitated by the development of RAD-seq assembly pipelines targeted at phylogenetic research (Eaton, 2014).

Incongruence between datasets has been a long-standing occurrence in molecular phylogenetic inference, traditionally manifesting as incongruences between different gene trees. The advance of next-generation sequencing (NGS) technology has shown that the issue is not solved by merely incorporating more data (Jeffroy et al., 2006). There are three possible categories

of confounding information in a phylogenetic study: noise, systematic error, and an underlying biological signal. Noise is an effect of the inherently stochastic nature of sequence evolution and leads to a deterioration of phylogenetic signal over time. As such, noise most heavily impacts very small datasets and deep nodes (Misof et al., 2014). Incongruence may also reflect a true biological signal, for example, the presence of ILS or non-tree-like evolution, i.e. introgression, hybridization, or recombination (Misof et al., 2014; Salichos et al., 2014). Effects of hybridization range from introgression of individual alleles, to organelle capture, to hybrid speciation (Currat et al., 2008; Stegemann et al., 2012; Twyford and Ennos, 2012). Either of these processes will result in discordant gene trees, and several approaches have been proposed to unravel them. Based on the distributions of conflicting phylogenetic patterns in the genome, it is possible to distinguish the more stochastic signal of ILS from the directional and asymmetric signal of hybridization (Durand et al., 2011).

Here, we apply RAD-seq to Hawaiian *Melicope*, a lineage with a crown age of ca. 10 Myr (Appelhans et al., 2018a). We use RAD-seq to infer species-level relationships in the lineage, in a phylogenetic context of several colonization events of individual islands, multiple possible bottlenecks, and adaptive radiations within a lineage. The taxonomic implications of our phylogenetic results are discussed within the framework of evidence for both ancient and current introgression.

MATERIALS AND METHODS

Taxon Sampling

Table 1 details the identity and origin of the 101 samples of this study: 6 outgroup and 95 ingroup specimens representing 41 Hawaiian species (81% of the lineage). Two samples represent the two independent colonization events to the Marquesas Islands (28% of Marquesan species). Taxonomic treatment follows species recognized in Wood et al. (2016) plus a recently described species (Wood et al., 2017) and including *Platydesma* (Appelhans et al., 2017). Additionally, morphologically divergent specimens of *Melicope barbigera* (KW16722 and KW16718) and *Melicope ovata* (KW16762, KW17082, and MA663) were included (**Table 1**, asterisk) to elucidate whether these might represent separate taxa. We also included two specimens, KW17111 and KW15733, which correspond closely, though not entirely, to the description of *Melicope wawraeana* as delimited by Stone et al. (1999). Even the O'ahu populations that were considered the core of *M. wawraeana* are variable, suggesting that it is a potentially artificial taxon (Stone et al., 1999). Since the morphology of the two specimens did not correspond entirely to the O'ahu populations considered to be *M. wawraeana*, we included them here as *Melicope* sp. (**Table 1**).

RAD Library Preparation

DNA was extracted from silica-dried material using the Qiagen DNeasy Plant Mini Kit® (Qiagen, Hilden, Germany) as per the manufacturer's instructions with incubation in lysis buffer elongated to 2h. DNA concentration was measured using the Qubit® fluorometer and the Qubit® dsDNA BR Assay Kit (Thermo Fisher Scientific, Darmstadt, Germany) and adjusted to 30 ng/μL. Floragenex Inc. (Portland, Oregon, USA) generated RAD libraries using the restriction enzyme *Sfb*I. With a method following Baird et al. (2008) being employed, including the use of sample-specific barcodes, the samples were sequenced on an Illumina® GAIIx platform to produce 100-bp single-end reads.

RAD Locus Assembly

Quality of raw reads was checked using FastQC (Andrews, 2010). The program *ipyrad* v.0.7.21 was used to demultiplex raw reads allowing a mismatch of 1 bp. Raw reads were trimmed using cutadapt v.1.9.1 (Martin, 2011) as implemented in *ipyrad* by removing adapter sequences, trimming bases with Phred scores <30 and removing reads shorter than 35 bp after trimming. Trimmed reads were assembled *de novo* using the *ipyrad* pipeline. The software attempts to evaluate orthology by scoring alignments of reads or sequences, as opposed to assessing purely sequence identity (Eaton, 2014). The alignment score is the user-determined clustering threshold to be met. To reduce the risk of introducing assembly error to our dataset, we performed a modified clustering optimization approach (Paris et al., 2017). We iterated over core clustering parameters and plotted assembly matrices (cluster depth, heterozygosity, number of putatively paralogous loci, number of single-nucleotide polymorphisms (SNPs)) to identify parameters introducing excessive assembly errors (Paetzold et al., unpublished results; Paris et al., 2017). In addition, we optimized the clustering of reads within each individual sample and the clustering of consensus sequences across loci separately, reasoning that the divergence found within each individual genome might be significantly different from the ca. 10 Myr of divergence (Appelhans et al., 2018a) within the lineage as a whole. Thus, the assembly was generated using a clustering threshold of 95 for in-sample clustering and 90 for between-sample clustering. The final filtering of loci was performed for values 10, 32, 50, 67, and 85 as the minimum numbers of samples per locus.

Phylogenetic Inference and Quartet Sampling

Phylogenetic inference was performed on all resulting alignments using maximum likelihood (ML) and Bayesian inference (BI). As individual loci are very short and may comprise a high fraction of missing data, a partitioned analysis is neither computationally feasible nor expected to produce reliable results. Thus, all datasets were analyzed solely concatenated. ML was performed using ExaML v3.0.2 (Kozlov et al., 2015) using the new rapid hill-climbing algorithm, a random number seed, the gamma model of rate heterogeneity, and the median for discrete approximation of rate heterogeneity. For datasets containing minimum numbers of 10, 32, and 50 samples, the memory saving option for gappy alignments was activated (-S). Parsimony starting trees were generated using RAxML v8.2.4 (Stamatakis, 2014). RAxML was also used to generate 100 bootstrap replicate alignments and their corresponding parsimony starting trees. ExaML searches were run on every replicate alignment with the above-mentioned settings.

BI was performed using ExaBayes v 1.5 (Aberer et al., 2014). Four independent runs were carried out with a convergence stopping criterion (split frequencies average <5% in three

TABLE 1 | Samples within this study including origin, voucher placement, and assignment to Stone's sections.

Species	Stone's section	Collection number, Herbarium voucher	Origin
***Melicope adscendens* (H. St. John & E. P. Hume) T. G. Hartley & B. C. Stone**	***Apocarpa***	**Appelhans MA628 (silica sample only, ORPF)**	**Maui**
Melicope anisata (H. Mann) T. G. Hartley & B. C. Stone	*Cubicarpa*	Appelhans MA665 (GOET, PTBG)	Kauaʻi
M. anisata (H. Mann) T. G. Hartley & B. C. Stone	*Cubicarpa*	Appelhans MA668 (GOET, PTBG, USA)	Kauaʻi
Melicope balloui (Rock) T. G. Hartley & B. C. Stone	*Megacarpa*	Wood KW7685 (PTBG)	Maui
***Melicope barbigera* A. Gray**	***Apocarpa***	**Appelhans MA666 (BISH, GOET, PTBG, USA)**	**Kauaʻi**
M. barbigera A. Gray	*Apocarpa*	Wood KW15333 (PTBG)	Kauaʻi
M. barbigera A. Gray	*Apocarpa*	Wood KW15449 (PTBG)	Kauaʻi
M. barbigera A. Gray	*Apocarpa*	Wood KW15961 (PTBG)	Kauaʻi
*M. barbigera** A. Gray	*Apocarpa*	Wood KW16722 (PTBG)	Kauaʻi
*M. barbigera** A. Gray	*Apocarpa*	Wood KW16718 (PTBG)	Kauaʻi
Melicope christophersenii (H. St. John) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA618 (BISH, GOET, PTBG, USA)	Oʻahu
Melicope christophersenii (H. St. John) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA621 (silica sample only, cultivated at Puʻu Kaʻala)	Oʻahu
Melicope clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA615 (GOET, PTBG)	Oʻahu
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA617	Oʻahu
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA634 (PTBG)	Maui
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA650 (GOET, PTBG, USA)	Maui
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA651 (BISH, GOET, PTBG, USA)	Maui
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA655 (silica sample only)	Maui
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA657 (GOET, PTBG, USA)	Maui
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA670	Kauaʻi
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA672	Kauaʻi
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA693	Hawaiʻi
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA695	Hawaiʻi
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Oppenheimer s.n. (silica sample only)	Maui
***M. clusiifolia* (A. Gray) T. G. Hartley & B. C. Stone**	***Pelea***	**Oppenheimer H91641 (US)**	**Lānaʻi**
M. clusiifolia (A. Gray) T. G. Hartley & B. C. Stone	*Pelea*	Wood KW16146 (PTBG)	Kauaʻi
***M. clusiifolia* (Gray) T. G. Hartley & B. C. Stone**	***Pelea***	**Appelhans MA675**	**Kauaʻi**
Melicope cornuta (Hillebr.) Appelhans, K. R. Wood & W. L. Wagner	*Platydesma*	Ching s.n. (silica sample only)	Oʻahu
***M. cornuta* var. *decurrens* (B. C. Stone) Appelhans, K. R. Wood & W. L. Wagner**	***Platydesma***	**Takahama s.n. (silica sample only)**	**Oʻahu**
Melicope cruciata (A. Heller) T. G. Hartley & B. C. Stone	*Megacarpa*	Wood KW16251 (PTBG)	Kauaʻi
Melicope degeneri (B. C. Stone) T. G. Hartley & B. C. Stone	*Cubicarpa*	Wood KW15903 (PTBG)	Kauaʻi
M. degeneri (B. C. Stone) T. G. Hartley & B. C. Stone	*Cubicarpa*	Wood KW15984 (PTBG)	Kauaʻi
Melicope feddei (H. Lév.) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA688 (BISH, GOET, PTBG, USA)	Kauaʻi
M. feddei (H. Lév.) T. G. Hartley & B. C. Stone	*Megacarpa*	Wood KW15844 (PTBG)	Kauaʻi
***M. haleakalae* (B. C. Stone) T. G. Hartley & B. C. Stone**	***Pelea***	**Appelhans MA645 (BISH, GOET, PTBG)**	**Maui**
M. haleakalae (B. C. Stone) T. G. Hartley & B. C. Stone	*Pelea*	Appelhans MA646 (BISH, GOET, PTBG, USA)	Maui
***Melicope haupuensis* (H. St. John) T. G. Hartley & B. C. Stone**	***Apocarpa***	**Appelhans MA687 (BISH)**	**Kauaʻi**
M. haupuensis (H. St. John) T. G. Hartley & B. C. Stone	*Apocarpa*	Wood KW16791 (PTBG)	Kauaʻi
M. haupuensis (H. St. John) T. G. Hartley & B. C. Stone	*Apocarpa*	Wood KW16794 (PTBG)	Kauaʻi
Melicope hawaiensis (Wawra) T. G. Hartley & B. C. Stone	*Apocarpa*	Appelhans MA633 (BISH, GOET, PTBG, USA)	Maui
M. hawaiensis (Wawra) T. G. Hartley & B. C. Stone	*Apocarpa*	Appelhans MA700	Hawaiʻi
M. hawaiensis (Wawra) T. G. Hartley & B. C. Stone	*Apocarpa*	Oppenheimer s.n. (silica sample only)	Maui
Melicope hiiakae (B. C. Stone) T. G. Hartley & B. C. Stone	*Megacarpa*	Ching s.n. (silica sample only)	Oʻahu
Melicope hivaoaensis J. Florence		Meyer 826	Hivaoa, Marquesas Islands
Melicope inopinata J. Florence		Meyer 887	Hivaoa, Marquesas Islands
***Melicope kavaiensis* (H. Mann) T. G. Hartley & B. C. Stone**	***Megacarpa***	**Appelhans MA679 (BISH, GOET, PTBG, USA)**	**Kauaʻi**

(Continued)

TABLE 1 | Continued

Species	Stone's section	Collection number, Herbarium voucher	Origin
Melicope knudsenii (Hillebr.) T. G. Hartley & B. C. Stone	*Apocarpa*	Appelhans MA629 (silica sample only, ORPF)	Maui
***M. knudsenii* (Hillebr.) T. G. Hartley & B. C. Stone**	***Apocarpa***	**Oppenheimer H41610 (BISH)**	**Maui**
M. knudsenii (Hillebr.) T. G. Hartley & B. C. Stone	*Apocarpa*	Wood KW17119 (PTBG)	Kaua'i
Melicope lydgatei (Hillebr.) T. G. Hartley & B. C. Stone	*Megacarpa*	Ching s.n. (silica sample only)	O'ahu
***Melicope makahae* (B. C. Stone) T. G. Hartley & B. C. Stone**	***Apocarpa***	**Takahama s.n. (silica sample only)**	**O'ahu**
M. makahae (B. C. Stone) T. G. Hartley & B. C. Stone (cf.)	*Apocarpa*	Appelhans MA609 (GOET, PTBG)	O'ahu
Melicope molokaiensis (Hillebr.) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA635 (BISH, GOET, PTBG)	Maui
M. molokaiensis (Hillebr.) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA643 (BISH, GOET, PTBG, USA)	Maui
***M. molokaiensis* (Hillebr.) T. G. Hartley & B. C. Stone**	***Megacarpa***	**Oppenheimer s.n. (silica sample only)**	**Maui**
Melicope mucronulata (H. St. John) T. G. Hartley & B. C. Stone	*Apocarpa*	Appelhans MA630 (silica sample only, ORPF)	Maui
Melicope munroi (St. John) T. G. Hartley & B. C. Stone	*Megacarpa*	Oppenheimer s.n. (silica sample only)	Lana'i
Melicope oahuensis (H. Lév.) T. G. Hartley & B. C. Stone	*Cubicarpa*	Appelhans MA610 (BISH, GOET, PTBG, USA)	O'ahu
M. oahuensis (H. Lév.) T. G. Hartley & B. C. Stone	*Cubicarpa*	Ching s.n. (silica sample only)	O'ahu
***Melicope oppenheimeri* K. R. Wood, Appelhans & W. L. Wagner**	***Megacarpa***	**Wood KW7419 (PTBG)**	**Maui**
M. oppenheimeri K. R. Wood, Appelhans & W. L. Wagner	*Megacarpa*	Wood KW7408 (PTBG)	Maui
Melicope orbicularis (Hillebr.) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA656 (BISH, GOET, PTBG, USA)	Maui
M. orbicularis (Hillebr.) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA659 (GOET, PTBG)	Maui
Melicope ovalis (St. John) T. G. Hartley & B. C. Stone	*Cubicarpa*	Wood KW13724 (PTBG)	Maui
***Melicope ovata* (H. St. John & E. P. Hume) T. G. Hartley & B. C. Stone**	***Apocarpa***	**Appelhans MA662 (GOET, PTBG, USA)**	**Kaua'i**
M. ovata (H. St. John & E. P. Hume) T. G. Hartley & B. C. Stone	*Apocarpa*	Appelhans MA684 (BISH, GOET)	Kaua'i
*M. ovata** (H. St. John & E. P. Hume) T. G. Hartley & B. C. Stone	*Apocarpa*	Appelhans MA663 (BISH, GOET, PTBG, USA)	Kaua'i
*M. ovata** (H. St. John & E. P. Hume) T. G. Hartley & B. C. Stone	*Apocarpa*	Wood KW17082 (PTBG)	Kaua'i
*M. ovata** (H. St. John & E. P. Hume) T. G. Hartley & B. C. Stone	*Apocarpa*	Wood KW16762 (PTBG)	Kaua'i
Melicope pallida (Hillebr.) T. G. Hartley & B. C. Stone	*Apocarpa*	Appelhans MA689 (silica sample only)	Kaua'i
M. pallida (Hillebr.) T. G. Hartley & B. C. Stone	*Apocarpa*	Wood KW16789 (PTBG)	Kaua'i
M. pallida (Hillebr.) T. G. Hartley & B. C. Stone	*Apocarpa*	Wood KW15571 (PTBG)	Kaua'i
***Melicope paniculata* (H. St. John) T. G. Hartley & B. C. Stone**	***Cubicarpa***	**Perlman 19387 (PTBG) = Appelhans MA660 (silica sample)**	**Kaua'i**
M. paniculata (H. St. John) T. G. Hartley & B. C. Stone	*Cubicarpa*	Wood KW16155 (PTBG)	Kaua'i
Melicope peduncularis (H. Lév.) T. G. Hartley & B. C. Stone	*Cubicarpa*	Appelhans MA652 (BISH, GOET, PTBG, USA)	Maui
M. peduncularis (H. Lév.) T. G. Hartley & B. C. Stone	*Cubicarpa*	Appelhans MA653 (BISH, GOET, PTBG, USA)	Maui
***Melicope pseudoanisata* (Rock) T. G. Hartley & B. C. Stone**	***Megacarpa***	**Appelhans MA632 (silica sample only, ORPF)**	**Maui**
M. pseudoanisata (Rock) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA636 (silica sample only)	Maui
M. pseudoanisata (Rock) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA642 (GOET, PTBG, USA)	Maui
Melicope puberula (H. St. John) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA680 (GOET, PTBG, USA)	Kaua'i
M. puberula (H. St. John) T. G. Hartley & B. C. Stone	*Megacarpa*	Wood KW16058 (PTBG)	Kaua'i
Melicope radiata (H. St. John) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA696	Hawai'i
***Melicope rostrata* (Hillebr.) Appelhans, K. R. Wood & W. L. Wagner**	***Platydesma***	**Appelhans MA683 (BISH, GOET)**	**Kaua'i**
***Melicope rotundifolia* (A. Gray) T. G. Hartley & B. C. Stone**	***Megacarpa***	**Ching s.n. (silica sample only)**	**O'ahu**

(Continued)

TABLE 1 | Continued

Species	Stone's section	Collection number, Herbarium voucher	Origin
Melicope sandwicensis (Hook. & Arn.) T. G. Hartley & B. C. Stone	*Apocarpa*	Ching s.n. (silica sample only)	O'ahu
Melicope sessilis (H. Lév.) T. G. Hartley & B. C. Stone	*Megacarpa*	Appelhans MA644 (BISH, GOET, PTBG, USA)	Maui
Melicope sp. (Rock) T. G. Hartley & B. C. Stone	*Megacarpa*	Wood KW17111 (PTBG)	Kaua'i
Melicope sp. (Rock) T. G. Hartley & B. C. Stone	*Megacarpa*	Wood KW15733 (PTBG)	Kaua'i
Melicope spathulata A. Gray	*Platydesma*	Appelhans MA697	Hawai'i
M. spathulata A. Gray	*Platydesma*	Wood KW16743 (PTBG)	Kaua'i
M. spathulata A. Gray	*Platydesma*	Wood KW16836 (PTBG)	Kaua'i
Melicope stonei K. R. Wood, Appelhans & W. L. Wagner	*Apocarpa*	Appelhans MA691	Kaua'i
M. stonei K. R. Wood, Appelhans & W. L. Wagner	*Apocarpa*	Wood KW16727 (PTBG)	Kaua'i
Melicope volcanica (A. Gray) T. G. Hartley & B. C. Stone (cf.)	*Megacarpa*	Oppenheimer s.n. (silica sample only)	Lāna'i
Melicope waialealae (Wawra) T. G. Hartley `& B. C. Stone	*Pelea*	Wood KW16015 (PTBG)	Kaua'i
Outgroup			
Melicope aneura (Lauterb.) T. G. Hartley		Appelhans MA418 (LAE, USA)	Papua New Guinea
Melicope durifolia (K. Schum.) T. G. Hartley		Appelhans MA455 (LAE, USA)	Papua New Guinea
Melicope polyadenia Merr. & L. M. Perry		Appelhans MA438 (LAE, USA)	Papua New Guinea
Melicope triphylla Merr.		Appelhans MA394 (GOET)	cultivated Hortus Botanicus Leiden
Melicope brassii T. G. Hartley		Appelhans MA436 (LAE, USA)	Papua New Guinea
M. durifolia (K. Schum.) T. G. Hartley		Appelhans MA465 (LAE, USA)	Papua New Guinea

Asterisk marks morphologically deviating specimens. Samples in bold were used in parameter optimization. ORPF, cultivated at Olinda Rare Plant Facility.

subsequent generations) and for a minimum of 100,000 generations sampling every 100th generation under the GTR+I+G model. Majority rule consensus trees were drawn on topologies of all four runs combined after the first 25% was discarded as burn-in.

Analysis of large-scale, concatenated datasets can result in erroneous relationships with high bootstrap support because of a failure to model the effects of ILS (Gadagkar et al., 2005; Kubatko and Degnan, 2007; Seo, 2008). These effects can be driven by only a few loci (Shen et al., 2017) and especially pertain to short branches (Kumar et al., 2012). On the other hand, a simulation study has shown that concatenated analysis of datasets containing loci with anomalous gene trees will more likely result in unresolved species tree topologies, rather than highly supported false ones (Huang and Knowles, 2009).

Methods implementing the multispecies coalescent (MSC) model explicitly incorporate gene tree conflict into species tree inference and are thus more robust to ILS than concatenation approaches (Kubatko and Degnan, 2007) but are often intractable for large datasets (Liu et al., 2015). Summary methods of species tree inference under the MSC, for example, ASTRAL (Mirarab et al., 2014) or NJst (Liu and Yu, 2011), are based on the analysis of individual gene trees and have become popular due to their comparative speed and accuracy. However, the limited information content of individual RAD loci often limits their application for gene tree inference, which may negatively impact species tree estimation (Salichos and Rokas, 2013; Mirarab et al., 2016). Alternatively, site-based methods avoid estimation of gene trees, instead using SNP data directly, and so are expected to be well suited to short, low-variability loci (Molloy and Warnow, 2018). We employed the SVDQuartets method, which infers quartet trees from SNPs using phylogenetic invariant patterns under the coalescent model and then infers the species tree by quartet joining of the subtrees using algebraic statistics (Chifman and Kubatko, 2014). We converted the SNP datasets into nexus format using the Ruby script *convert_vcf_to_nexus.rb* (Matschiner, 2018). The SVDQuartets analysis was computed as implemented in the software PAUP*4.0a (Swofford, 2002; Swofford, 2018). We analyzed 250,000 randomly selected quartets and assessed statistical support using 100 nonparametric bootstrap support replicates. For ambiguous positions in the SNP matrix, we chose the "Distribute" option, as these positions represent heterozygous sites.

To estimate the robustness of resolved relationships, we employed the Quartet Sampling method, which aims to measure branch support in large sparse alignments (Pease et al., 2018). As each internal branch divides all samples within a phylogeny into four non-overlapping subsets, the method randomly samples one taxon per subset to produce a quartet phylogeny. The topology of each quartet is either concordant with the tree topology or discordant. Discord is measured and quantified to produce four metrics—quartet concordance (QC), quartet differential (QD), quartet informativeness (QI), and quartet fidelity (QF)—allowing effective assessment of branch-related (QC, QD, and QI) and taxon-related (QF) discordance in the dataset (Pease et al., 2018). The method is implemented in the python script *quartet_sampling.py* (https://www.github.com/fephyfofum/quartetsampling). We performed Quartet Sampling on all datasets and the respectively resolved topologies using 500 replicates per branch with a minimum required overlap of 300,000 bp in the min10, min32, min50, and min67 concatenated datasets. The minimum overlap was lowered to 140,000 bp in the

min85 concatenated dataset, as otherwise five samples would have been excluded from the analysis.

Test for Introgression

The *D*-statistics (Durand et al., 2011) is a site-based test for introgression. In a four-taxon topology (((P1, P2), P3), O), a derived allele in the P3 lineage is expected to occur also in either P1 or P2 with equal frequency, giving rise to either an ABBA or BABA discordant site pattern (Durand et al., 2011). A statistically significant imbalance in these site pattern frequencies provides evidence of introgression, while equal frequencies are associated with neutral processes like ILS. Unfortunately, this test is not well suited for deeper evolutionary timescales, where the P3 lineage has diverged into multiple sub-lineages, and it also does not allow inference of direction of introgression. Partitioned *D*-statistics is a system of multiple four-taxon *D*-statistics in a symmetric, five-taxon phylogeny with the ingroup taxa forming two pairs (P1, P2) and ($P3_1$, $P3_2$) and an outgroup taxon (O) (**Figure 1**) (Eaton and Ree, 2013). The partitioned *D*-statistics identifies sites, in which either or both of the P3 lineages share a derived allele with either P1 or P2, but not both (**Figure 1**) (Eaton and Ree, 2013).

We used partitioned *D*-statistics to infer whether discordant relationships inferred between major clades (see below) are caused by ILS or introgression. We defined entire clades as lineages and tested all combinations obeying the symmetric topology.

RESULTS

Raw Data and Assembly

Illumina Sequencing yielded an average of 10,439,082 reads per sample (342,914–34,663,109). After quality trimming, an average of 10,327,562 reads per sample (271,257–34,542,777) were left.

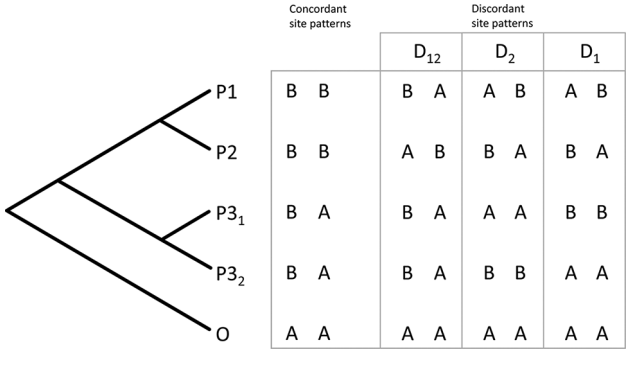

FIGURE 1 | The principle of five-taxon *D*-statistics test. Biallelic site patterns are quantified, which support or contradict the underlying symmetric phylogeny. Asymmetry of discordant site patterns is quantified to calculate three separate *D*-statistics characterizing introgression from the $P3_1$ taxon (D_1), the $P3_2$ taxon (D_2), or their common ancestor (D_{12}) into the taxa designated P_1 and P_2 (Eaton and Ree, 2013).

The assembled dataset contained a total of 786,169 clusters prior to filtering by sample coverage. Filtering reduced the number of loci by over 90% (**Table 2**). The final datasets contained between 7,266 (min85) and 59,041 (min10) loci. The number of variable sites (SNPs) ranged from 529,045 (min10) to 82,760 (min85) (**Table 2**).

Phylogenetic Inference

All five final datasets were used for phylogenetic inference in concatenated BI, ML, and SVD Quartets analyses. Statistical support for inferred relationships was assessed using posterior probabilities (PPs), nonparametric bootstrap (NBS) (ML-NBS and SVD-NBS), and Quartet Sampling. Analyses of the five datasets resulted in mostly congruent relationships, with few exceptions (see below). NBS and PP values are very high across the trees. QI values are high for all nodes (>0.9), and QF scores are average between 0.83 and 0.88 across datasets. **Figure 2** shows the result of phylogenetic inference in the concatenated min32 dataset.

Hawaiian *Melicope* are divided into five main clades corresponding to those previously resolved by Appelhans et al. (2014b). These five clades are fully supported by all statistical methods. The former genus *Platydesma* represents the earliest diverging lineage (clade V; **Figure 2**). Clade IV corresponds to Stone's section *Pelea*, characterized by whorled leaves. The remaining Stone sections appear to be non-monophyletic. Species ascribed to Stone's section *Apocarpa* are resolved as two independent lineages (Clades II and III). Clade I comprises all species of Stone's sections *Cubicarpa* and *Megacarpa* intermingled (**Figure 2**). Relationships of clade III were resolved incongruently between datasets and analyses. BI and ML analyses resolved clade III as sister to clade IV, and the resulting monophyletic lineage again in a sister–group relationship to clades I + II with maximum PP and high ML-NBS support in four of the datasets (min10, min32, min50, and min85), yet with some discord detected by Quartet Sampling (**Figures 2** and **3**, **Supplemental Figures 1**, **2**, and **4**). The concatenated min67 dataset resolves clade III as sister to clades I + II, and clade IV as sister to clades I + II + III (**Supplemental Figure 3**) with medium statistical support. Coalescent-based SVDQuartets analysis of SNP datasets resolved a third alternative topology. Here, clade II is resolved as sister to clade III, and the resulting lineage is sister to clades I + IV. This topology receives medium-to-low SVD-NBS support across all SNP datasets, as well as medium-to-high negative QC values, indicating substantial counter-support for this relationship (**Supplemental Figures 5–9**). The relationship of clade III is highly discordant over quartet replicates (**Supplemental Figure 3**). Across all datasets, the discord detected by Quartet Sampling for the ancestral branch is skewed favoring one of the tested alternative quartet topologies (QD; **Figure 3**, **Supplemental Figures 1–9**).

The remaining relationships within individual clades are fully resolved, improving resolution to the species and intraspecies levels (**Figure 2**). The majority of all Hawaiian *Melicope* are resolved in clade I, and relationships among species show many nodes with notable discord and very short branches (**Figure 2**). Most of the nodes show low QC and medium-to-low QD values (**Figure 3**). Three samples show incongruent relationships between datasets. This pertains to the Marquesan *Melicope hivaoaensis*, which is resolved in clade I as either sister to the remaining species (**Supplemental Figures 3–9**) within the clade or diverging prior to *Melicope lydgatei* (**Figures 2**

TABLE 2 | Differences between the number of loci, their concatenated length, and the number of SNPs resulting from filtering by minimum samples per locus (10, 32, 50, 67, and 85).

	Total	min10	min32	min50	min67	min85
Number of loci	786,169	59,041	36,622	30,801	23,401	7,266
Concatenated length (bp)	NA	4,800,367	2,986,760	2,506,242	1,892,473	584,086
Number of SNPs	NA	529,045	385,871	332,935	256,276	82,760

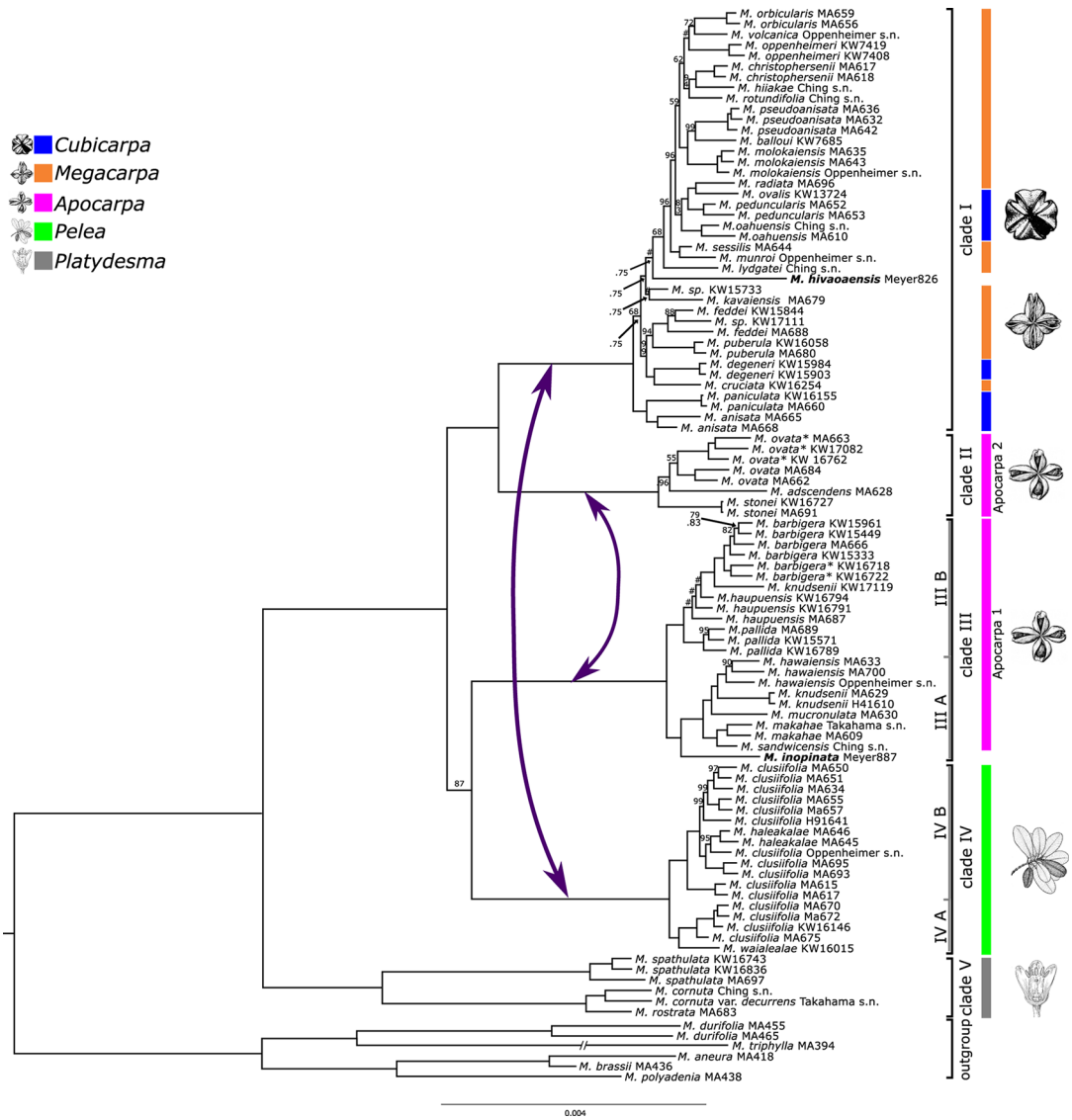

FIGURE 2 | Phylogeny of Hawaiian *Melicope* based on the concatenated min32 dataset. Bayesian posterior probability (PP) values are indicated above branches, and maximum likelihood (ML) nonparametric bootstrap support (ML-NBS) below branches. Support values are not shown for maximally supported clades (1.00pp/100BS). A hashtag (#) represents incongruent species relationships between Bayesian and ML analyses. Clade colors and line drawings correspond to morphologically limited Stone's sections. Bold samples represent Marquesan species. Asterisks mark specimens differing morphologically from the typical representatives of these species. Purple arrows mark putative introgression events.

and **3**, **Supplemental Figures 1** and **2**) as well as to *M. kavaiensis* and *Melicope* sp. KW15773 (**Figure 3**, **Supplemental Figures 1–9**). In all datasets, QC values show high discord or even counter-support for the placement of these three specimens. However, while QD and QF values are high for *M. hivaoaensis*, for both *M. kavaiensis* and *Melicope* sp. KW15773, QD values are low and QF scores are below average (0.47–0.6 for *M. kavaiensis*) (**Figure 3**, **Supplemental Figures 1–9**). The remaining relationships in clade I are congruent among all concatenation-based analyses. Site-specific coalescence analysis, however, resolved largely incongruent relationships for taxa

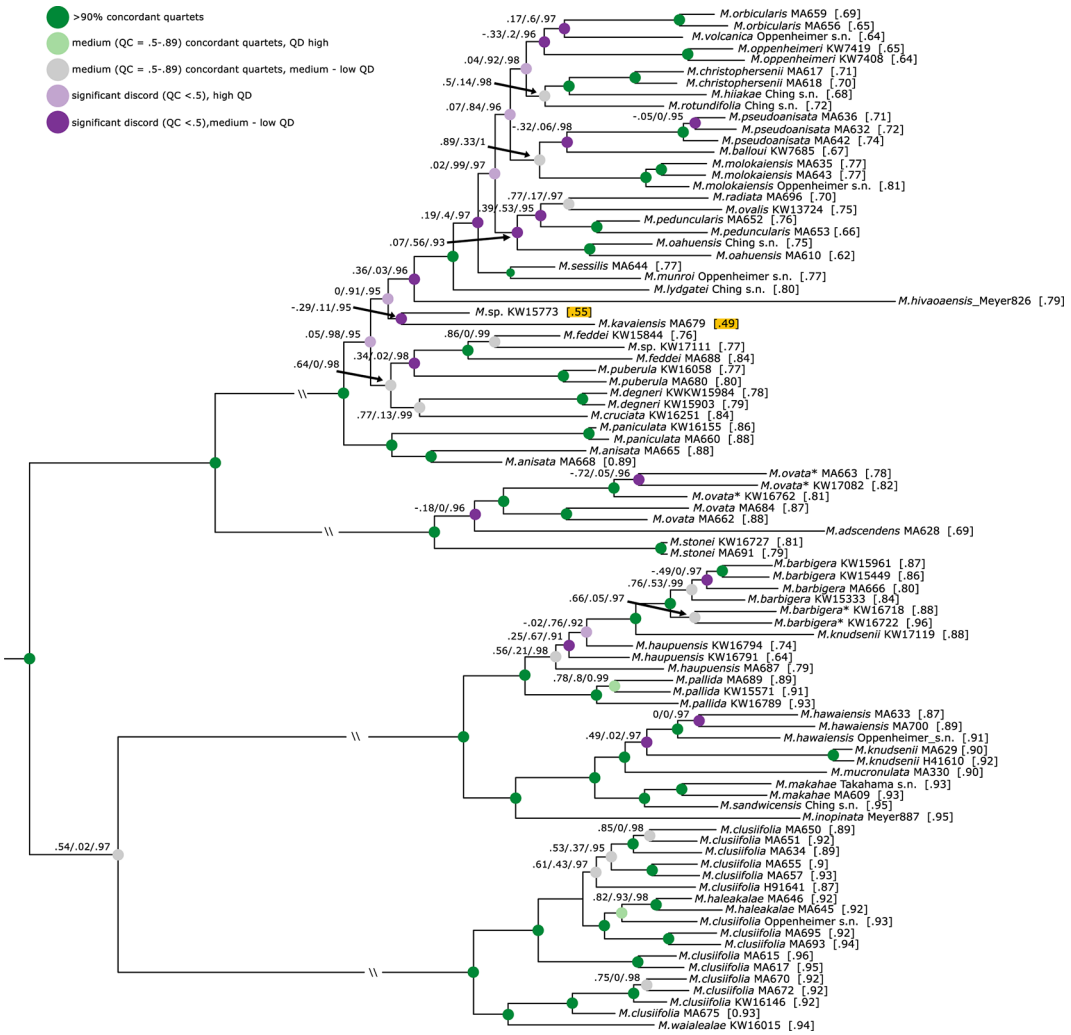

FIGURE 3 | Phylogeny of Hawaiian *Melicope* based on the min32 dataset. Quartet Sampling results (quartet concordance (QC)/quartet differential (QD)/quartet informativeness (QI)) are indicated on branches, and quartet fidelity (QF) values behind samples. Nodes are colored according to QC and QD values. Results are not shown for branches with QC > 0.9. The lowest QF values are highlighted. Outgroup specimens are removed for graphical purposes. All outgroup relationships receive maximum QC values (1/-/1).

in this clade, especially pertaining to the most recent divergences. The inferred relationships receive medium-to-very-low SVD-NBS values and show a high amount of discord in Quartet Sampling (**Supplemental Figures 5–9**).

Clades III and IV are subdivided into two subclades each. Most species sampled with multiple accessions are resolved as monophyletic with high support and no discord detected in Quartet Sampling. Exceptions are *Melicope clusiifolia*, *Melicope haupuensis*, *Melicope knudsenii*, and *Melicope feddei*. *M. clusiifolia* is resolved paraphyletic with respect to *Melicope haleakalae*, which is nested within clade IVB with high-to-maximum support. Specimens of *M. haupuensis* are resolved as polyphyletic within clade IIIB. The relationships among the three sampled taxa are not resolved consistently across datasets and poorly supported. Quartet Sampling reveals a high level of discord and below-average QF scores (**Figure 3**, **Supplemental Figures 1–9**).

M. knudsenii is also resolved as polyphyletic with two Maui specimens (MA629 and H41610) monophyletic in clade IIIA, while the third sample from Kaua'i (KW17119) is resolved as sister to *M. barbigera* in clade IIIB (**Figure 2**). Either relationship is virtually uncontested (**Figures 2** and **3**, **Supplemental Figures 1–9**). *M. feddei* is paraphyletic with respect to one of the Kaua'i *M. wawraeana*-like specimens (KW17111). The three individuals form a fully supported, monophyletic unit (**Figures 2** and **3**, **Supplemental Figures 1–9**).

None of the three species complexes (*M. elliptica*, *M. kavaiensis*, and *M. volcanica* complexes) are resolved as monophyletic. Species of both the *M. kavaiensis* and *M. volcanica* complexes are resolved in clade I (**Figure 2**) in proximity to each other, but not sister to each other. Species of the *M. elliptica* complex are resolved in different subclades of clade III (**Figure 2**). Both *M. barbigera* and *M. ovata* were resolved as

monophyletic, and the morphologically divergent specimens (**Table 1**, asterisk) are resolved as sister clades to the samples with the typical morphology of the respective species with high support (**Figure 2**).

The species from the Marquesas Islands are deeply nested within the Hawaiian clade. *Melicope inopinata* is resolved in clade III as sister to the rest of subclade IIIA. *M. hivaoaensis* represents a group of six morphologically similar species that form a highly supported monophyletic clade (Appelhans et al., 2014b; Appelhans et al., 2018a) and is nested within clade I here (**Figure 2**).

Test for Introgression

The min32 dataset was used for the ABBA-BABA test, since it produced the highest number of fully supported nodes. The tree topology in **Figure 2** was chosen to represent the species tree topology, as it was recovered by the majority of analyses. The *D*-statistics was only used to test the incongruent position of clade III, as for incongruent species within clade I, the sampling of the respective populations is not sufficient to draw reliable conclusions. Samples within clades were pooled, and SNP frequencies were used for *D*-statistic calculations (Durand et al., 2011). All possible relationships complying with the *D*-statistic assumptions were tested. A total of 24,673 loci covered at least one-third of all samples per clade and, thus, contributed to the test results. **Table 3** summarizes the tested topologies and inferred partitioned *D*-statistics. When clades III and IV are tested as donors for introgressed loci, values for D_{12} are small and not significant ($Z_{12} < 2.55$), while values for D_1 and D_2 are significant, respectively. For tests with either of clade I or II designated as P3 lineages, D_{12}, and D_1 and D_2, are all significant (**Table 3**). For all tested configurations, the dataset exhibits more than 3,000 discordant site patterns (**Table 3**).

TABLE 3 | Partitioned *D*-statistics for introgression involving clades I–IV.

((P1, P2), (P3$_1$, P3$_2$), O)	D_{12}	Z_{12}	*n* ABAAA	*n* BABBA
((I, II), (III, IV), V&O)	0.020	0.95	809.84	778.1
((I, II), (IV, III), V&O)	–0.020	0.96	778.1	809.84
((IV, III), (I, II), V&O)	0.066	3.28	1,273.58	1,115.49
((IV, III), (II, I), V&O)	–0.066	3.29	1,273.58	1,115.5
((P1, P2), (P3$_1$, P3$_2$), O)	D_1	Z_1	*n* ABBAA	*n* BABAA
((I, II), (III, IV), V&O)	0.276	8.48	261.01	437.67
((I, II), (IV, III), V&O)	–0.276	8.17	437.67	261.01
((IV, III), (I, II), V&O)	–0.242	7.07	403.07	222.05
((IV, III), (II, I), V&O)	0.290	7.61	271.16	444.45
((P1, P2), (P3$_1$, P3$_2$), O)	D_2	Z_2	*n* ABABA	*n* BAABA
((I, II), (III, IV), V&O)	–0.253	7.08	505.48	286.73
((I, II), (IV, III), V&O)	0.253	7.05	286.73	505.48
((IV, III), (I, II), V&O)	0.290	7.57	271.16	444.45
((IV, III), (II, I), V&O)	–0.242	7.24	403.06	222.05

Z scores ≥2.55 represent a significant value for D_x. The respective numbers of concordant and discordant site patterns are listed. Clade numbers refer to those in ***Figure 2****, and the group they are assigned to in the partitioned D-statistics test is indicated (compare* ***Figure 1****). O, outgroup.*

DISCUSSION

Phylogeny and Introgression

Analysis of *ipyrad* assemblies consistently resolved five major clades within Hawaiian *Melicope* (**Figure 2**). However, the relationships of clade III were incongruent among the five datasets and analysis methods (**Figure 2**, **Supplemental Figures 1–9**). Incongruence between datasets may be caused by one of three factors: noise, ILS, or non-tree-like evolution. As noise is expected to impact small datasets and deep nodes most severely (Misof et al., 2014), it is unlikely a sufficient cause of the incongruence observed here, since our RAD-seq alignments are substantial in size (**Table 1**) and the remaining deep nodes are not affected.

The QD values of the branch illustrate that one of the discordant topologies is inferred significantly more often (0.0–0.4; **Figure 3**, **Supplemental Figures 1–9**), which indicates non-tree-like evolution as the cause for the discord. Thus, we used the partitioned *D*-statistics to test for signals of ancient introgression between clades I through IV with all clades tested as putative donor (P3) lineages. In all cases, values for D_1 and D_2 were each significant, yet values for D_{12} were only significant when clades I and II were defined as P3 (**Table 3**). Positive values of D_1 represent introgression between P2 and P3$_1$, while negative values indicate introgression between P1 and P3$_1$, and values for D_2 represent events analogous for P3$_2$ and P2 (Eaton and Ree, 2013; Pease and Hahn, 2015). The significant values for D_1 and D_2 indicate introgression between the respective ancestors of clades I and IV as well as between respective ancestors of clades II and III. Significant values for D_{12} represent shared ancestral alleles from the clade I + II progenitor introduced into the respective ancestor of clades III and IV (**Figure 2**, **Table 3**). All taxa in clades II and III have apocarpous fruits, while all taxa in clades I and IV have syncarpous fruits (Stone et al., 1999), providing a morphological connection between either of the two pairs, which might be linked to introgressed information. However, we interpret these result cautiously, as *D*-statistic results are sensitive to confounding signals from multiple introgressive events due to phylogenetic non-independence of tests (Eaton et al., 2015).

The origin of the Hawaiian *Melicope* lineage predates the rise of the current high islands (Appelhans et al., 2018a). Thus, the inferred introgressive events are associated with a time when the ancestral species were still relegated either to refugial areas on small, low islands or shortly after they colonized the young island of Kaua'i. The time frame under consideration presents a "bottleneck" scenario, where the ancestral lineages were likely in close spatial proximity. Additionally, increased volcanic activity of the Hawaiian hot spot coincided with the rise of Kaua'i (Price and Clague, 2002). This volcanic activity could have produced lava flows, earthquakes, tsunamis, and other catastrophic events, which may have additionally promoted hybridization (Stuessy et al., 2014). The ancestral hybridization events may even have promoted subsequent adaptive radiation on the islands (Kagawa and Takimoto, 2018). Estimation of divergence times in Hawaiian *Melicope* will be needed to infer the time frame for hybridization events in ancestral lineages. While there is strong evidence for ancient hybridization events within Hawaiian *Melicope*, the nature of *de novo* RAD-seq data currently limits our analytic

methods. Further information may be obtained through gene tree-based approaches applied to target capture or whole genome-sequencing data (Meng and Kubatko, 2009) or by examining SNP-based patterns, as they vary spatially along a reference genome (Martin et al., 2013).

Bootstrap and PP support values were generally high across trees inferred from different datasets but generally increased with dataset size. Lenient filtering in RAD-seq data is often practiced, as there is a correlation between the size of a data matrix and resolution and support of relationships (Wagner et al., 2013; Hodel et al., 2017). RAD locus dropout is expected to increase with increasing divergence times, as enzyme cut sites will be lost or gained through mutation (Cariou et al., 2013). Loci with a small amount of missing data are therefore expected to represent the conserved spectrum of genomic sites and, thus, provide a limited capacity of resolution. On the other hand, sparse loci are expected to iincrease resolution of relationships despite also introducing noise, as they are assumed to represent the more rapidly evolving genomic fractions (Cariou et al., 2013; Wagner et al., 2013; Eaton et al., 2015). However, including all loci is not advisable either, as there seems to be a point at which inclusion of increasingly more sparse loci might start to decrease support. At this point, noise, due to missing data introduced by the inclusion of more sparse loci, will overpower the informative value these loci provide. However, the Quartet Sampling method seems an adequate approach to evaluating the reliability of the dataset, as the QC value showed the same trend in all datasets regardless of size and offer the QI score to assess the amount and impact of missing data.

We detected some discord between relationships resolved by concatenation and site-specific coalescence-based methods (**Figure 2**, **Supplemental Figures 1–9**). The evaluation of the performance of different species-tree inference methods is a matter of ongoing research, especially with regard to genomic datasets. Concatenation-based ML inference can be statistically inconsistent under some conditions in the MSC, that is, ILS causing gene trees to differ from the true species tree (Kubatko and Degnan, 2007). However, the limits of the concatenated approach are poorly understood (Molloy and Warnow, 2018), and the performance of concatenated Bayesian analysis has yet to be formally assessed. Some simulation studies show that concatenated RAD-seq data are robust to gene tree/species tree discord when inferring relationships among taxa (Rivers et al., 2016). In addition, concatenated approaches potentially offer hidden support as a feature overriding gene tree/species conflict (Gatesy and Springer, 2014; Rivers et al., 2016), although hidden support has not been addressed in plant phylogenomic research yet. Coalescence-based methods are statistically consistent under the MSC. Bayesian co-estimation of gene trees and the species tree under the MSC is currently considered the most effective approach, yet computationally very demanding and thus less applicable to large datasets. Hence, summary and site-specific MSC methods have become popular, and several algorithms implementing the concepts do exist (Liu et al., 2015). However, the assessment of the performance of these methods under empirical and simulated conditions is still a matter of active research. For example, gene tree methods have proven to be statistically inconsistent if the cause of gene tree discord is horizontal gene transfer, instead of ILS (Solís-Lemus et al., 2016; Fernández-Mazuecos et al., 2018). Several recent simulation studies compared the accuracy of multiple summary and site-based coalescent methods, including SVDQuartets, as well as concatenated ML under varying levels of ILS and gene tree estimation error (GTEE). Concatenated ML was at least competitive with MSC methods under most conditions and outperformed SVDQuartets under all tested conditions, including high GTEE (Chou et al., 2015; Mirarab et al., 2016; Molloy and Warnow, 2018). The latter would be expected in RAD-seq datasets and should also be present herein.

With respect to species relationships inferred for Hawaiian *Melicope* and considering the observed lower accuracy of SVDQuartets compared with concatenation-based approaches under conditions typically characterizing RAD datasets, we suggest that the results from concatenated BI and ML are probably more accurate than those based on SVDQuartets and will be discussed below. However, we do stress that none of the approaches have proven to be statistically consistent under conditions observed herein, that is, ILS, GTEE, and horizontal gene transfer (**Figure 2**).

Taxonomic Implications

The former small genus *Platydesma* and Stone's section *Pelea* are each monophyletic (**Figure 2**), while the three remaining sections of Stone, comprising the majority of all Hawaiian *Melicope* species, are not. *Apocarpa* is divided into two lineages with the majority of species resolved in *Apocarpa* 1 (**Figure 2**). The three species of the *Apocarpa* 2 clade share a number of morphological traits, though none of them is either exclusive or inclusive. All species of *Apocarpa* 2 occur in mesic forests only and, with the exception of *Melicope stonei*, share a sprawling, shrubby habit (Stone et al., 1999; Wood et al., 2017). Finally, in all *Apocarpa* 2 species, both endocarp and exocarp are glabrous and inflorescences are few-flowered, though both of these traits also appear outside of this group (Stone et al., 1999; Wood et al., 2017). In a previous analysis, apocarpous species were resolved in three different clades (Appelhans et al., 2014b), one of which, consisting of *M. elliptica* only, could not be sampled in this study. Further research will be necessary to identify morphological character combinations distinguishing these lineages. Stone's sections *Cubicarpa* and *Megacarpa* are paraphyletic with respect to each other (**Figure 2**) with species of each resolved intermingled throughout the clade. The two groups differ by the degree of carpel connation, with carpels "connate from base up to 2/3 of their length" (Stone et al., 1999) characterizing *Megacarpa* and carpels "nearly to completely" connate (Stone et al., 1999) characterizing *Cubicarpa*. Carpel connation clearly represents a continuum and not two discrete units. As there is no pattern to the degree of carpel fusion apparent in clade I, the separation of these two of Stone's sections seems artificial.

Interspecies relationships within clade I are less well supported than in the remaining clades, and Quartet Sampling reveals measurable discord at nearly every branch in the backbone of this clade. For many of the nodes with low QC values, QD values are high (**Figure 3**, **Supplemental Figures 1–9**), which characterizes ILS and corresponds to the shortness of these branches. On the other hand, many branches show low QD values, indicating

widespread introgression between these lineages. Unfortunately, sampling herein is not sufficient to test individual relationships.

Of the 24 species represented by multiple accessions, 20 were resolved as monophyletic, while four species were either paraphyletic or polyphyletic. *M. clusiifolia* is the most widespread and morphologically diverse of all Hawaiian *Melicope* (Stone et al., 1999), and it is paraphyletic with both of the other species of Stone's section *Pelea*, *M. haleakalae* and *Melicope waialealae* (clade IV, **Figure 2**). Several attempts have been made to subdivide *M. clusiifolia* into varying constellations of subspecies, varieties, and forms (St. John, 1944; Stone, 1969). In the most recent taxonomic treatment, Stone et al. (1999) synonymized all subdivisions of the species, arguing that the variable characters seem to represent a continuum rather than distinguishable, discrete units. However, the authors also issued the recommendation that the overall pattern of variability in *M. clusiifolia* should be studied in detail (Stone et al., 1999). *M. haleakalae* is characterized as differing from *M. clusiifolia*, mainly in its persistent sepals (Stone et al., 1999). Considering that *M. haleakalae* is nested deeply within *M. clusiifolia* (**Figure 2**, clade IV), the two might be regarded as conspecific and included in an overall evaluation of the complex. *M. waialealae* differs from *M. clusiifolia* mainly in leaf shape (Stone et al., 1999). However, since the leaf shape of *M. clusiifolia* is highly variable, *M. waialealae* might represent one end of a continuum across both taxa rather than one of two distinct states. On the other hand, these three species might represent a case of speciation in progress. In this case, the deep nesting, especially of *M. haleakalae*, within *M. clusiifolia*, would represent speciation following a progenitor-derivative scenario (Crawford, 2010). The widespread, morphologically variable *M. clusiifolia* would meet all criteria of the progenitor (p) species. The persistent petals in *M. haleakalae* and the leaf shape in *M. waialealae* would represent a variable, morphological feature in the parent being fixed in the respective derivative (d) species. Identification of a true p–d relationship is difficult and rare. However, several candidate species pairs do exist (Crawford, 2010). The p–d species pair *Layia glandulosa* (Hook.) Hook. & Arn. and *Layia discoidea* D. D. Keck (Asteraceae) show not only a shift in morphology between progenitor and derivative species but also geographic isolation due to a shift in habitat (Baldwin, 2005). This could be the same for *M. waialealae*, which is restricted to bogs, whereas the putative progenitor *M. clusiifolia* occurs in mesic to wet forests (Stone et al., 1999). Unfortunately, there are no data available regarding breeding system or pollinator communities in these species, creating potential barriers to gene flow. Detailed studies of morphological characters, gene flow, and abiotic habitat factors are necessary to determine whether these taxa are separate p–d species pairs or conspecific, as already indicated in previous studies (Appelhans et al., 2014b).

M. knudsenii, delimited by Stone et al. (1999) as the only species occurring on non-adjacent islands, was resolved as polyphyletic, with three samples resolved as two distinct lineages within clade III. Appelhans et al. (2014b) already showed that this taxon is polyphyletic, consisting of three taxa. One of these was recently described as *M. stonei* (Wood et al., 2017). Our results confirm the previously resolved pattern with the two specimens of *M. knudsenii* from Maui resolved as sister to *Melicope hawaiensis* and the specimen from Kaua'i as sister to *M. barbigera* (clade III, **Figure 2**). We confirm that these specimens clearly represent different species. The Maui species will be resurrected under one of the names used in an earlier treatment by Stone (1969), wherein he adopted a narrower species concept than in the later classification (Stone et al., 1999), leaving *M. knudsenii* restricted to only populations on Kaua'i.

The three specimens of *M. haupuensis* included in this study are resolved as paraphyletic. Moreover, they are the only species resolved with incongruent topologies of the individual samples associated with the different datasets (compare **Figure 2**, **Supplemental Figures 1–9**). Quartet Sampling shows strong discord for either of the inferred relationships with medium QD values (**Figure 3**, **Supplemental Figures 1–9**), indicating the possibility of introgressed sites. Moreover, QF scores for the three specimens are considerably lower than the average, indicating a rogue behavior (Aberer et al., 2013; Pease et al., 2018) of the three taxa. Additionally, several specimens in the field were observed presenting morphologically intermediate forms between *M. haupuensis* and *M. barbigera* (personal observation K.R. Wood). QD values for the latter are also low (**Figure 3**, **Supplemental Figures 1–9**). Both the morphological intermediates and the incongruence associated with different datasets indicate potential hybridization between these species. However, conclusively identifying putative hybridization events would require sampling at the population level, including any morphological intermediates.

Multiple samples of *M. ovata* and *M. barbigera* were included in our analyses, representing both the typical morphology and a deviating morphotype. For either species, the morphologically deviating samples were resolved as the sister group to the samples with the typical habit. Variant morphotypes of *M. ovata* displayed a pubescent lower leaf surface, whereas leaves are typically glabrous in this species. *M. barbigera* usually has few-flowered inflorescences (Stone et al., 1999). In contrast, the variant morphotype has inflorescences with a considerably larger number of flowers. Genomic divergence is comparable with that of other species pairs within the lineage. Both groups might be another case of speciation in progress within Hawaiian *Melicope*. In both cases, detailed morphological studies will be necessary to investigate if the morphologically divergent populations of the two species should be recognized as separate taxa.

The two *M. wawraeana*-like specimens are resolved in clade I, but not closely related to each other. One specimen (KW17111) is nested within the two samples of *M. feddei* with high support (**Figures 2** and **3**). *M. wawraeana* is very similar to *M. feddei* and differs mainly in pedicel length (Stone et al., 1999). The present results suggest that some populations might be conspecific with *M. feddei*, while others (e.g., from the herein unsampled type location) are not. The relationships of the second *M. wawraeana*-like specimen (KW15733) are resolved incongruently among datasets, as are the relationships of the sampled specimen of *M. kavaiensis*. The two samples are resolved either as sister groups (**Figure 3**, **Supplemental Figures 1**, **4**, **5**, and **7–9**) or as consecutive sister clades within clade I (**Supplemental Figures 2**, **3**, and **6**). There is a substantial amount of discord in the dataset for either of the resolved relationships. QD values are low, indicating the possibility of introgression between these morphologically distinct species. Additionally, QF scores for either of the specimens are low corresponding to the rogue behavior of the samples.

The rogue behavior of the aforementioned samples (*M. kavaiensis*, *Melicope* sp. KW15733) might also be related to the incongruent placement of *M. hivaoaensis*, as the three taxa are inferred as closely related, regardless of the relation to the remainder of clade I. For this specimen, QC and QD values are low; however, QF is high (**Figure 3**, **Supplemental Figures 1–9**). *M. hivaoaensis* represents an adaptive radiation of five species endemic to the Marquesas Islands, whose predecessor colonized from the Hawaiian Islands (Appelhans et al., 2014a; Appelhans et al., 2018a). Successful island colonizations have been associated with recent hybridization or polyploidization events (Paetzold et al., 2018). There was no polyploidization event immediately prior to the colonization of the Hawaiian Islands itself (Paetzold et al., 2018), making a polyploidization event prior to the colonization of the Marquesas Islands unlikely. Chromosome counts for Marquesan species are not available for a conclusive answer. However, results herein indicate the presence of several hybridization events within the lineage. Thus, a hybridization event might have predated the colonization of the Marquesas Islands as well. As the incongruent position of *M. hivaoaensis* seems to correspond to the rogue behavior of *Melicope* sp. KW15733 and *M. kavaiensis*, the latter two might represent the parental lineages of the Marquesan *Melicope* radiation. A conclusive answer to the question is contingent on a thorough sampling of all concerned lineages as well as a prior revision of the *M. wawraeana* species concept.

We confirm previous results showing that Hawaiian *Melicope* colonized the Marquesas Islands twice independently, negating the hypothesis that the remote Hawaiian Islands constitute a dispersal sink (Harbaugh et al., 2009; Appelhans et al., 2014a; Appelhans et al., 2018a). The nesting of Marquesan species in different Hawaiian clades is corroborated by fruit morphology (Hartley, 2001), since *M. hivaoaensis* and its close relatives from the Marquesas Islands have syncarpous fruits as do the species in clade I, while *M. inopinata* has apocarpous fruits like the species in clade III.

The present study provides unprecedented insight into the relationships of Hawaiian *Melicope*. Several previous findings could be corroborated and firmly supported by genome-wide data, including the non-monophyly of most of Stone's sections, which cannot be held up as delimited (Stone, 1969; Stone et al., 1999). The lineage is in need of a taxonomic revision. Understanding the relationships of Hawaiian *Melicope* would be enhanced by some formal recognition of the subclades with corresponding morphological features. However, the creation of novel formal subgroups within *Melicope* section *Pelea* must also include the extra-Hawaiian members of the section. The former genus *Platydesma* is the most distinctive group within *Melicope* sect. *Pelea* and should receive some level of formal recognition. *Apocarpa* species need to be split into two groups, one of which would include the Marquesan species *M. inopinata*. However, conclusive treatment of *Apocarpa* should be adjourned until an improved understanding of the separation within the *M. elliptica* complex is attained. Delimitations of species within the *Pelea* group, *M. barbigera*, *M. ovata*, and *M. haupuensis*, may need revision, but levels of hybridization should also be investigated as part of that process. *M. wawraeana* requires revision as well as a prerequisite to test the putative hybrid character of the Marquesan radiation. Furthermore, the other six *Melicope* species endemic to the Marquesas Islands would need to be included in a novel taxonomic recognition of Stone's former sections *Megacarpa* and *Cubicarpa*.

DATA AVAILABILITY

All demultiplexed raw read data were submitted to the NCBI Sequence Read Archive; BioProject number PRJNA559258.

AUTHOR CONTRIBUTIONS

MA, CP, and WW conceived and designed the study. MA, CP, and KW collected the samples. CP carried out the laboratory work and performed all analyses. DE provided valuable input for the analyses. CP drafted the manuscript, and all authors contributed to writing and editing. All authors have read and approved the final manuscript.

FUNDING

This project was financially supported by the German Science Foundation (DFG; Grant AP 251/3-1 to MA). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We are grateful to Charmian Dang for support with collection permits and to H. Alves, R. Belcher, S. Ching, K. Fay, K. Kosaka, S. Marquez, H. Oppenheimer, S. Perlman, J. Price, K. Range, T. Takahama, K. Togikawa, and A. Williams for help in collecting the specimens. We thank the United States Department of Land and Natural Resources (Permits: P-242, KPI2017-102, ODF-051316R, and MDF-092216A) for the permission to collect plants in forest reserves on Kaua'i, O'ahu, Maui, and Hawai'i (Big Island); the Nature Reserve for the permission to collect plants at the Waikamoi Preserve on Maui; the Puu Kukui Watershed Preserve for the permission to collect along the Puu Kukui Trail; and the United States Fish & Wildlife Service for the permission to export samples (Permit: MA96221B-O). We also thank Alice Tangerini for some pencil drawings of Stone's sections of *Melicope*. The remaining drawings were reproduced from Otto Degener's Flora Hawaiiensis (1960–1970). We thank two anonymous reviewers and Gonzalo N. Feliner for critical reading and suggestions that helped improved an earlier version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01074/full#supplementary-material

REFERENCES

Aberer, A. J., Kobert, K., and Stamatakis, A. (2014). ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. *Mol. Biol. Evol.* 31, 2553–2556. doi: 10.1093/molbev/msu236

Aberer, A. J., Krompass, D., and Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst. Biol.* 62, 162–166. doi: 10.1093/sysbio/sys078

Andrews, S. (2010). FastQC: a quality control tool from high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc

Appelhans, M. S., Reichelt, N., Groppo, M., Paetzold, C., and Wen, J. (2018b). Phylogeny and biogeography of the pantropical genus *Zanthoxylum* and its closest relatives in the proto-Rutaceae group (Rutaceae). *Mol. Phylogenet. Evol.* 126, 31–44. doi: 10.1016/j.ympev.2018.04.013

Appelhans, M. S., Wen, J., Duretto, M., Crayn, D., and Wagner, W. L. (2018a). Historical biogeography of *Melicope* (Rutaceae) and its close relatives with a special emphasis on Pacific dispersals. *J. Syst. Evol.* 56, 576–599. doi: 10.1111/jse.12299

Appelhans, M. S., Wen, J., and Wagner, W. L. (2014a). A molecular phylogeny of *Acronychia*, *Euodia*, *Melicope* and relatives (Rutaceae) reveals polyphyletic genera and key innovations for species richness. *Mol. Phylogenet. Evol.* 79, 54–68. doi: 10.1016/j.ympev.2014.06.014

Appelhans, M. S., Wen, J., Wood, K. R., Allan, G. J., Zimmer, E. A., and Wagner, W. L. (2014b). Molecular phylogenetic analysis of Hawaiian Rutaceae (*Melicope*, *Platydesma* and *Zanthoxylum*) and their different colonization patterns: molecular phylogeny of Hawaiian Rutaceae. *Bot. J. Lin. Soc.* 174, 425–448. doi: 10.1111/boj.12123

Appelhans, M. S., Wood, K. R., and Wagner, W. L. (2017). Reduction of the Hawaiian genus *Platydesma* into *Melicope* section *Pelea* (Rutaceae) and notes on the monophyly of the section. *PhytoKeys* 91, 125–137. doi: 10.3897/phytokeys.91.21363

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3, e3376. doi: 10.1371/journal.pone.0003376

Baldwin, B. G. (2005). Origin of the serpentine-edemic herb *Layia discoidea* from the widespread *L. glandulosa* Compositae. *Evolution* 59, 2473–2479. doi: 10.1111/j.0014-3820.2005.tb00956.x

Baldwin, B. G., and Sanderson, M. J. (1998). Age and rate of diversification of the Hawaiian silversword alliance (Compositae). *Proc. Natl. Acad. Sci. U.S.A.* 95, 9402–9406. doi: 10.1073/pnas.95.16.9402

Barrier, M., Baldwin, B. G., Robichaux, R. H., and Purugganan, M. D. (1999). Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the Hawaiian silversword alliance (Asteraceae) inferred from floral homeotic gene duplications. *Mol. Biol. Evol.* 16, 1105–1113. doi: 10.1093/oxfordjournals.molbev.a026200

Cariou, M., Duret, L., and Charlat, S. (2013). Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecol. Evol.* 3, 846–852. doi: 10.1002/ece3.512

Carlquist, S. (1967). The biota of long-distance dispersal. V. Plant dispersal to Pacific Islands. *Bull. Torrey Bot. Club* 94, 129–162. doi: 10.2307/2484044

Carlquist, S. (1974). *Island biology*. New York, NY: Columbia University Press.

Carr, G. D. (1998). "Chromosome evolution and speciation in Hawaiian flowering plants," in *Evolution and speciation of island plants*. Eds. T. F. Stuessy and M. Ono (Cambridge, UK: Cambridge University Press), 97–119.

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

Crawford, D. J. (2010). Progenitor-derivative species pairs and plant speciation. *TAXON* 59, 1413–1423. doi: 10.1002/tax.595008

Currat, M., Ruedi, M., Petit, R. J., and Excoffier, L. (2008). The hidden side of invasions: massive introgression by local genes. *Evolution* 62, 1908–1920. doi: 10.1111/j.1558-5646.2008.00413.x

Díaz-Arce, N., Arrizabalaga, H., Murua, H., Irigoien, X., and Rodríguez-Ezpeleta, N. (2016). RAD-seq derived genome-wide nuclear markers resolve the phylogeny of tunas. *Mol. Phylogenet. Evol.* 102, 202–207. doi: 10.1016/j.ympev.2016.06.002

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* 28, 2239–2252. doi: 10.1093/molbev/msr048

Eaton, D. A. R. (2014). PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics* 30, 1844–1849. doi: 10.1093/bioinformatics/btu121

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

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

Eaton, D. A. R., Spriggs, E. L., Park, B., and Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Syst. Biol.* 66, 399–412. doi: 10.1093/sysbio/syw092

Eggens, F., Popp, M., Nepokroeff, M., Wagner, W. L., and Oxelman, B. (2007). The origin and number of introductions of the Hawaiian endemic *Silene* species (Caryophyllaceae). *Am. J. Bot.* 94, 210–218. doi: 10.3732/ajb.94.2.210

Emerson, B. C. (2002). Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. *Mol. Ecol.* 16, 951–966.

Fernández-Mazuecos, M., Mellers, G., Vigalondo, B., Sáez, L., Vargas, P., and Glover, B. J. (2018). Resolving recent plant radiations: power and robustness of genotyping-by-sequencing. *Syst. Biol.* 67, 250–268. doi: 10.1093/sysbio/syx062

Francisco-Ortega, J., Santos-Guerra, A., Kim, S.-C., and Crawford, D. J. (2000). Plant genetic diversity in the Canary Islands: a conservation perspective. *Am. J. Bot.* 87, 909–919. doi: 10.2307/2656988

Gadagkar, S. R., Rosenberg, M. S., and Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Exp. Zool. B Mol. Dev. Evol.* 304B, 64–74. doi: 10.1002/jez.b.21026

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

Givnish, T. J. (1998). "Adaptive radiation and molecular systematics: issues and approaches," in *Molecular evolution and adaptive radiation*. Eds. T. J. Givnish and K. J. Sytsma (Cambridge, UK: Cambridge University Press), 1–54.

Givnish, T. J., Millam, K. C., Mast, A. R., Paterson, T. B., Theim, T. J., Hipp, A. L., et al. (2009). Origin, adaptive radiation and diversification of the Hawaiian lobeliads (Asterales: Campanulaceae). *Proc. R. Soc. Lond. [Biol.]* 276, 407–416. doi: 10.1098/rspb.2008.1204

Harbaugh, D. T. (2008). Polyploid and hybrid origins of Pacific Island sandalwoods (*Santalum*, Santalaceae) inferred from low-copy nuclear and flow cytometry data. *Int. J. Plant. Sci.* 169, 677–685. doi: 10.1086/533610

Harbaugh, D. T., Wagner, W. L., Allan, G. J., and Zimmer, E. A. (2009). The Hawaiian Archipelago is a stepping stone for dispersal in the Pacific: an example from the plant genus *Melicope* (Rutaceae). *J. Biogeogr.* 36, 230–241. doi: 10.1111/j.1365-2699.2008.02008.x

Hartley, T. G. (2001). On the taxonomy and biogeography of *Euodia* and *Melicope* (Rutaceae). *Allertonia* 8, 1–328.

Hipp, A. L., Eaton, D. A. R., Cavender-Bares, J., Fitzek, E., Nipper, R., and Manos, P. S. (2014). A framework phylogeny of the American oak clade based on sequenced RAD data. *PLoS One* 9, e93975. doi: 10.1371/journal.pone.0093975

Hodel, R. G. J., Chen, S., Payton, A. C., McDaniel, S. F., Soltis, P., and Soltis, D. E. (2017). Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering. *Sci. Rep.* 7. doi: 10.1038/s41598-017-16810-7

Huang, H., and Knowles, L. L. (2009). What is the danger of the anomaly zone for empirical phylogenetics? *Syst. Biol.* 58, 527–536. doi: 10.1093/sysbio/syp047

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genet.* 22, 225–231. doi: 10.1016/j.tig.2006.02.003

Kagawa, K., and Takimoto, G. (2018). Hybridization can promote adaptive radiation by means of transgressive segregation. *Ecol. Lett.* 21, 264–274. doi: 10.1111/ele.12891

Keeley, S. C., and Funk, V. A. (2011). "Origin and evolution of Hawaiian endemics: new patterns revealed by molecular phylogenetic studies," in *The biology of island floras*. Eds. D. Bramwell and J. Caujape-Castells (Cambridge, UK: Cambridge University Press), 57–88. doi: 10.1017/CBO9780511844270.005

Kozlov, A. M., Aberer, A. J., and Stamatakis, A. (2015). ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* 31, 2577–2579. doi: 10.1093/bioinformatics/btv184

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

Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L., and Tamura, K. (2012). Statistics and truth in phylogenomics. *Mol. Biol. Evol.* 29, 457–472. doi: 10.1093/molbev/msr202

Lens, F., Davin, N., Smets, E., and del Arco, M. (2013). Insular woodiness on the Canary Islands: a remarkable case of convergent evolution. *Int. J. Plant. Sci.* 174, 992–1013. doi: 10.1086/670259

Liu, L., and Yu, L. (2011). Estimating species trees from unrooted gene trees. *Syst. Biol.* 60, 661–667. doi: 10.1093/sysbio/syr027

Liu, L., Wu, S., and Yu, L. (2015). Coalescent methods for estimating species trees from phylogenomic data. *J. Syst. Evol.* 53, 380–390. doi: 10.1111/jse.12160

Losos, J. B., and Ricklefs, R. E. (2009). Adaptation and diversification on islands. *Nature* 457, 830–836. doi: 10.1038/nature07893

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10–12. doi: 10.14806/ej.17.1.200

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

Matschiner, T. (2018). Convert_vcf_to_nexus.rb. GitHub repository, https://github.com/mmatschner/tutorials/species_tree_inference_with_snp_data/.

Meng, C., and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Popul. Biol.* 75, 35–45. doi: 10.1016/j.tpb.2008.10.004

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., and Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* 17, 240–248. doi: 10.1101/gr.5681207

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

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

Misof, B., Meusemann, K., von Reumont, B. M., Kück, P., Prohaska, S. J., and Stadler, P. F. (2014). A priori assessment of data quality in molecular phylogenetics. *Algorithms Mol. Biol.* 9. doi: 10.1186/s13015-014-0022-4

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

Nepokroeff, M., Sytsma, K. J., Wagner, W. L., and Zimmer, E. A. (2003). Reconstructing ancestral patterns of colonization and dispersal in the Hawaiian understory tree genus *Psychotria* (Rubiaceae): a comparison of parsimony and likelihood approaches. *Syst. Biol.* 52, 820–838. doi: 10.1080/10635150390251072

Paetzold, C., Kiehn, M., Wood, K. R., Wagner, W. L., and Appelhans, M. S. (2018). The odd one out or a hidden generalist: Hawaiian *Melicope* (Rutaceae) do not share traits associated with successful island colonization. *J. Syst. Evol.* 56, 621–636. doi: 10.1111/jse.12454

Paris, J. R., Stevens, J. R., and Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods Ecol. Evol.* 8, 1360–1373. doi: 10.1111/2041-210X.12775

Pease, J. B., and Hahn, M. W. (2015). Detection and Polarization of introgression in a five-taxon phylogeny. *Syst. Biol.* 64, 651–662. doi: 10.1093/sysbio/syv023

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

Price, J. P., and Clague, D. A. (2002). How old is the Hawaiian biota? Geology and phylogeny suggest recent divergence. *Proc. R. Soc. Lond. [Biol.]* 269, 2429–2435. doi: 10.1098/rspb.2002.2175

Price, J. P., and Wagner, W. L. (2004). Speciation in Hawaiian angiosperm lineages: cause, consequence, and mode. *Evolution* 58, 2185–2200. doi: 10.1111/j.0014-3820.2004.tb01597.x

Price, J. P., and Wagner, W. L. (2018). Origins of the Hawaiian flora: phylogenies and biogeography reveal patterns of long-distance dispersal. *J. Syst. Evol.* 56, 600–620. doi: 10.1111/jse.12465

Ree, R. H., and Hipp, A. L. (2015). "Inferring phylogenetic history from restriction site associated DNA (RADseq)," in *Next-generation sequencing in plant systematics*. Eds. E. Hörandl and M. S. Appelhans (Königstein, Germany: Koeltz Scientific Books), 181–204.

Rivers, D. M., Darwell, C. T., and Althoff, D. M. (2016). Phylogenetic analysis of RAD-seq data: examining the influence of gene genealogy conflict on analysis of concatenated data. *Cladistics* 32, 672–681. doi: 10.1111/cla.12149

Roderick, G. K. (1997). Herbivorous insects and the Hawaiian silversword alliance: coevolution or cospeciation? Available at: http://scholarspace.manoa.hawaii.edu/handle/10125/3219 (Accessed March 14, 2019).

Roy, T., Cole, L. W., Chang, T.-H., and Lindqvist, C. (2015). Untangling reticulate evolutionary relationships among New World and Hawaiian mints (Stachydeae, Lamiaceae). *Mol. Phylogenet. Evol.* 89, 46–62. doi: 10.1016/j.ympev.2015.03.023

Rubin, B. E. R., Ree, R. H., and Moreau, C. S. (2012). Inferring phylogenies from RAD sequence data. *PLoS One* 7, e33394. doi: 10.1371/journal.pone.0033394

Salichos, L., and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 327–331. doi: 10.1038/nature12130

Salichos, L., Stamatakis, A., and Rokas, A. (2014). Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.* 31, 1261–1271. doi: 10.1093/molbev/msu061

Seo, T.-K. (2008). Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol. Biol. Evol.* 25, 960–971. doi: 10.1093/molbev/msn043

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

Solís-Lemus, C., Yang, M., and Ané, C. (2016). Inconsistency of species tree methods under gene flow. *Syst. Biol.* 65, 843–851. doi: 10.1093/sysbio/syw030

St. John, H. (1944). Diagnoses of Hawaiian species of *Pelea* (Rutaceae)—Hawaiian plant studies, 13. *Lloydia* 7, 265–274.

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. *PNAS* 109, 2434–2438. doi: 10.1073/pnas.1114076109

Stone, B. C. (1969). *The genus Pelea A. Gray. (Rutaceae, Evodiinae.) A taxonomic monograph*. 3rd ed. Stuttgart, Germany: J. Cramer.

Stone, B. C., Wagner, W. L., and Herbst, D. R. (1999). "Rutaceae," in *Manual of the flowering plants of Hawai'i*, Revised Edition. Eds. W. L. Wagner and S. H. Sohmer (Honolulu, HI, USA: University of Hawaii Press and Bishop Museum Press), 1174–1216.

Stuessy, T. F., Takayama, K., López-Sepúlveda, P., and Crawford, D. J. (2014). Interpretation of patterns of genetic variation in endemic plant species of oceanic islands: genetic variation in island plants. *Bot. J. Lin. Soc.* 174, 276–288. doi: 10.1111/boj.12088

Swofford, D. L. (2002). *Phylogenetic analysis using parsimony (* and other methods)*. Version 4.0b10. Sunderland, MA, USA: Sinauer Associates.

Swofford, D. L. (2018) PAUP:* Phylogenetic analysis using parsimony (and other methods). version 4a165. Available at: https://people.sc.fsu.edu/~dswofford/paut_test/.

Twyford, A. D., and Ennos, R. A. (2012). Next-generation hybridization and introgression. *Heredity* 108, 179–189. doi: 10.1038/hdy.2011.68

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., et al. (2013). Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.* 22, 787–798. doi: 10.1111/mec.12023

Wood, K. R., Appelhans, M. S., and Wagner, W. L. (2016). *Melicope oppenheimeri*, section Pelea (Rutaceae), a new species from West Maui, Hawaiian Islands: with notes on its ecology, conservation, and phylogenetic placement. *PhytoKeys* 69, 51–64. doi: 10.3897/phytokeys.69.8844

Wood, K. R., Appelhans, M. S., and Wagner, W. L. (2017). *Melicope stonei*, section Pelea (Rutaceae), a new species from Kaua'i, Hawaiian Islands: with notes on its distribution, ecology, conservation status, and phylogenetic placement. *PhytoKeys* 83, 119–132. doi: 10.3897/phytokeys.83.13442

Wu, Z.-Y., Monro, A. K., Milne, R. I., Wang, H., Yi, T.-S., Liu, J., et al. (2013). Molecular phylogeny of the nettle family (Urticaceae) inferred from multiple loci of three genomes and extensive generic sampling. *Mol. Phylogenet. Evol.* 69, 814–827. doi: 10.1016/j.ympev.2013.06.022

Conflict of Interest Statement: 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: 17 October 2019
doi: 10.3389/fpls.2019.01224

Phylogenomics Yields New Insight Into Relationships Within Vernonieae (Asteraceae)

Carolina M. Siniscalchi[1,2], Benoit Loeuille[3], Vicki A. Funk[4], Jennifer R. Mandel[1] and José R. Pirani[2]*

[1] The Mandel Lab, Department of Biological Sciences, University of Memphis, Memphis, TN, United States, [2] Laboratório de Sistemática Vegetal, Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil, [3] Departamento de Botânica - CCB, Universidade Federal de Pernambuco, Recife, Brazil, [4] Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington, DC, United States

Asteraceae, or the sunflower family, is the largest family of flowering plants and is usually considered difficult to work with, not only due to its size, but also because of the abundant cases of polyploidy and ancient whole-genome duplications. Traditional molecular systematics studies were often impaired by the low levels of variation found in chloroplast markers and the high paralogy of traditional nuclear markers like ITS. Next-generation sequencing and novel phylogenomics methods, such as target capture and Hyb-Seq, have provided new ways of studying the phylogeny of the family with great success. While the resolution of the backbone of the family is in progress with some results already published, smaller studies focusing on internal clades of the phylogeny are important to increase sampling and allow morphological, biogeography, and diversification analyses, as well as serving as basis to test the current infrafamilial classification. Vernonieae is one of the largest tribes in the family, accounting for approximately 1,500 species. From the 1970s to the 1990s, the tribe went through several reappraisals, mainly due to the splitting of the mega genus *Vernonia* into several smaller segregates. Only three phylogenetic studies focusing on the Vernonieae have been published to date, both using a few molecular markers, overall presenting low resolution and support in deepest nodes, and presenting conflicting topologies when compared. In this study, we present the first attempt at studying the phylogeny of Vernonieae using phylogenomics. Even though our sampling includes only around 4% of the diversity of the tribe, we achieved complete resolution of the phylogeny with high support recovering approximately 700 nuclear markers obtained through target capture. We also analyzed the effect of missing data using two different matrices with different number of markers and the difference between concatenated and gene tree analysis.

Keywords: Compositae, Hyb-Seq, phylogeny, target capture, *Vernonia*

OPEN ACCESS

Edited by:
Lisa Pokorny, National Institute of Agricultural and Food Research and Technology, Spain

Reviewed by:
Luis Palazzesi, National Council for Scientific and Technical Research (CONICET), Argentina
Sònia Garcia, Spanish National Research Council (CSIC), Spain

****Correspondence:***
Carolina M. Siniscalchi
carol.siniscalchi@gmail.com

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

Received: *04 June 2019*
Accepted: *04 September 2019*
Published: *17 October 2019*

Citation:
Siniscalchi CM, Loeuille B, Funk VA, Mandel JR and Pirani JR (2019) Phylogenomics Yields New Insight Into Relationships Within Vernonieae (Asteraceae). Front. Plant Sci. 10:1224. doi: 10.3389/fpls.2019.01224

INTRODUCTION

The Asteraceae, or Sunflower family, comprise about 10% of the diversity of angiosperms and are widespread occurring in almost all biomes and environments. Some groups comprise major components in threatened ecosystems, like the tribes Lychnophorinae in the Brazilian *campos rupestres* and Corymbieae and Arctotideae in the South African *fynbos* (Karis et al., 2009; Loeuille

et al., 2019). More than 40 species have been domesticated, e.g., lettuce, artichoke, sunflower, safflower, stevia, and chicory, and some noxious weeds also belong in the family, e.g., *Mikania micrantha* Kunth., *Chromolaena odorata* (L.) R.M.King & H.Rob., and *Ambrosia artemisiifolia* L. (ragweed).

Although the systematics of the family has been studied since before the Linnean system (e.g., de Tournefort, 1700; Vaillant, 1719-1723), and the most used infrafamiliar classification has remained largely unchanged since its publication (Cassini, 1819), our understanding of the phylogenetic relationships within the family has drastically changed in the last decades. Morphological and molecular phylogenies challenged the long-standing view that the Heliantheae alliance was the earliest diverging tribe in the family, showing that they are actually highly nested within the family (Jansen and Palmer, 1988; Bremer, 1994; Funk et al., 2009).

Nevertheless, tackling the backbone phylogeny of the family has always been challenging, as there is a well-documented evidence of an abundance of polyploidy, hybridization events, ancient whole-genome duplications, and explosive radiations (Barker et al., 2008; Semple and Watanabe, 2009; Barker et al., 2016). In the past 5 years, with the availability of second-generation sequencing methods and their adaptation for use with non-model organisms, two different approaches to understanding evolutionary relationships in Asteraceae have emerged. The first is the use of a set of RNA or DNA probes that target specific orthologous loci within the genome, allowing them to be captured, enriched, and sequenced (Mandel et al., 2014), and the second is the use of transcriptome sequencing to acquire orthologous loci (Huang et al., 2016), with both being used to produce family-level phylogenies (Huang et al., 2016; Mandel et al., 2017; Mandel et al., 2019).

While transcriptome sequencing is straightforward in relation to sample processing and wet lab procedures, the main drawbacks are the need to collect samples in a way that preserves the RNA in the tissue, which precludes using herbarium specimens as sources for sampling, and the fact that gene expression is variable, which may impact locus recovery across samples and making the possibility of combining data from different studies challenging (Wen et al., 2015).

Target capture associated with genome skimming arose initially as a way to obtain sequences from ultraconserved elements in the genome of vertebrates and invertebrates (Cronn et al., 2012; Faircloth et al., 2012; Grover et al., 2012) but has been further extended into plant phylogenomics recently, with the release of lower or higher taxonomic level probes, such as for family Asteraceae (Mandel et al., 2014), genera *Protea* (Mitchell et al., 2017), *Heuchera* (Folk et al., 2015), and *Inga* (Nicholls et al., 2015) and, more recently, for all angiosperms (Johnson et al., 2019). Although sample preparation requires extra steps, time, and additional cost from the target capture kit, target recovery is usually consistent within a lineage and allows the combination of data generated across different studies. Given the possibility of using previously collected material for DNA extraction, such as from herbarium collections, samples preserved in silica gel, or DNA banks, target capture is appealing in the context of the increasing challenges of securing financial and human resources for field work.

The Asteraceae conserved ortholog set (COS) kit developed by Mandel et al. (2014) has been successfully tested across the family (Mandel et al., 2017; Mandel et al., 2019) and within higher-nested lineages (Herrando-Moraira and The Cardueae Radiations Group, 2018). Aiming to study the effectiveness of this method in a lineage known for its complicated evolutionary and taxonomic history, we generated a phylogeny of tribe Vernonieae.

Vernonieae contains about 1,500 species and is distributed in the New and Old World, with the main diversity centers in Africa and South America. Members of Vernonieae are easily recognized by the homogamous heads composed only by tubular florets, the predominance of pinkish-purplish corollas, and the often recurved style branches (**Figure 1**). The circumscription of the tribe has hardly changed since Cassini's first description (1819), but genera circumscription within it has drastically changed since the 1980s.

Most of the species of the tribe have been previously placed in the comprehensive genus *Vernonia*, with more than 1,000 species. There were several attempts at creating infrageneric classifications for *Vernonia* (Jones, 1979; Jones, 1981; Jeffrey, 1988), which culminated in its pulverization into several other genera (Robinson, 1999a; Robinson, 1999b), such as *Centrapalus*, *Cyrtocymura*, *Distephanus*, *Lepidaploa*, *Lessingianthus*, and *Vernonanthura*.

The first phylogenetic studies in Vernonieae focused on the relationships within *Vernonia* (Keeley and Turner, 1990) and showed that the African species of the genus form a grade leading to a more speciose clade of New World species. In addition, their work demonstrated that the species now included in *Distephanus* (**Figure 1A**), a genus from Madagascar and South Africa, were the sister to the whole genus. Keeley et al. (2007) expanded this first phylogeny, using two chloroplast regions (*ndh*F and *trn*L-F) and ribosomal ITS, and focused on the whole tribe, already including several of the taxonomic changes that occurred since 1990 (**Figure 2A**). Again, the division between Old World and New World groups is clear, as well as the outgroup position of *Distephanus*. The complexity of the relationships in the New World clade also becomes evident with several instances where members of clades are found in distant locations, such as the clade formed by *Vernonia* s.str., found in North America that is a sister to a clade formed by genera from Central America and Brazil, which in turn is a sister to a large clade of Brazilian species. *Stokesia* (**Figure 1D**), a monotypic genus from the Southeastern USA and the only species in the tribe to present zygomorphic florets, seems somewhat problematic, with its position varying depending on the markers used but generally emerges close to Leiboldiinae, in the transition between the larger Old and New World clades in combined analyses.

In 2015, Loeuille et al. published an in-depth phylogeny of the American Vernonieae (**Figure 2B**), focusing on the evolution of secondary heads on the group, using internal transcribed spacer, two chloroplast regions (*ndh*F and *trn*L-F), and a morphological matrix. The division between Old and New World was also found, but *Distephanus* was not sampled. In this work, it was clear that some of the new subtribes and even new genera proposed in the years before were not monophyletic, such as subtribe Vernoniinae, whose members are spread out in several clades or

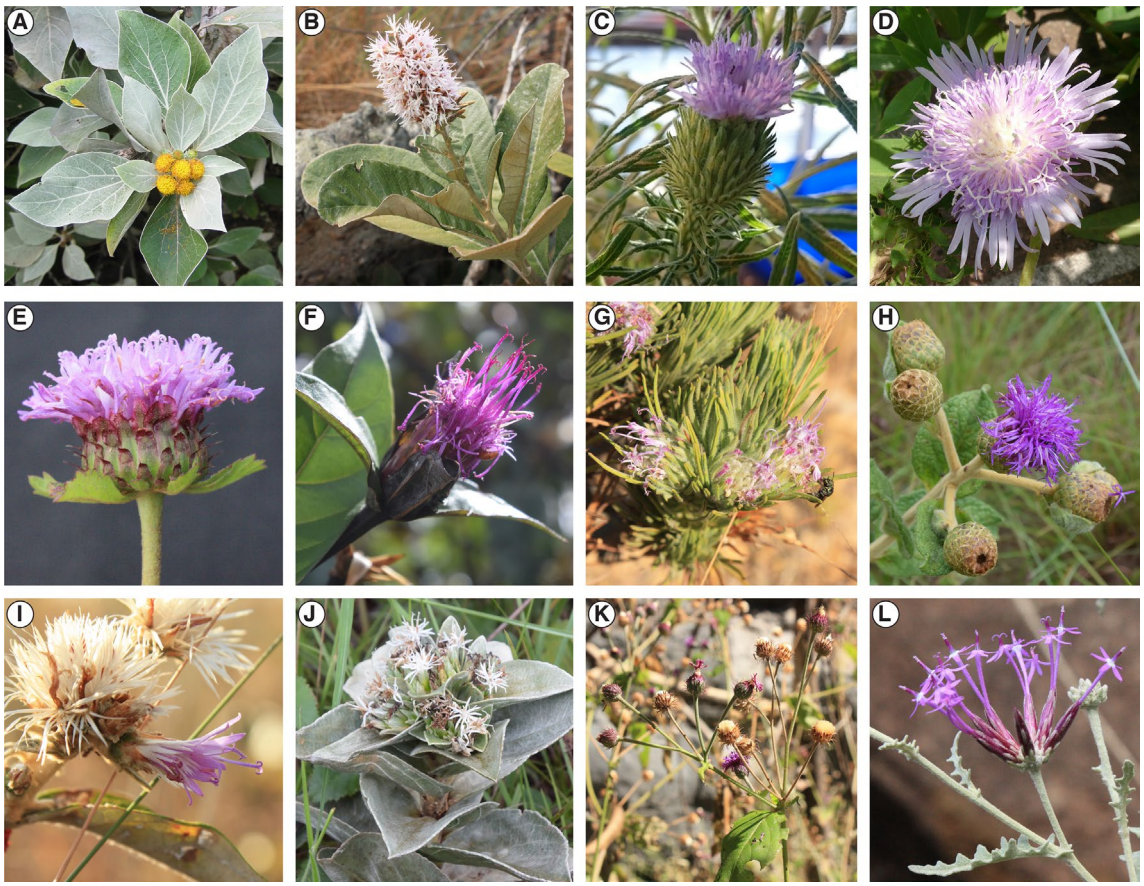

FIGURE 1 | Diversity in Vernonieae and related taxa. **(A)** *Distephanus populifolius* (Distephaninae). **(B)** *Moquinia racemosa* (tribe Moquinieae), **(C)** *Centrapalus pauciflorus* (Centrapalinae), **(D)** *Stokesia laevis* (Stokesiinae), **(E)** *Centratherum punctatum* (Lychnophorinae), **(F)** *Hololepis pedunculata* (Lychnophorinae), **(G)** *Lychnophora ericoides* (Lychnophorinae), **(H)** *Lessingianthus monocephalus* (Lepidaploinae), **(I)** *Strophopappus speciosus* (Lepidaploinae), **(J)** *Soaresia velutina* (Elephantopinae), **(K)** *Heterocypsela andersonii* (Dypterocypselinae), **(L)** *Chresta hatschbachii* (Chrestinae). Photos by VF **(A)** and CS **(B–L)**.

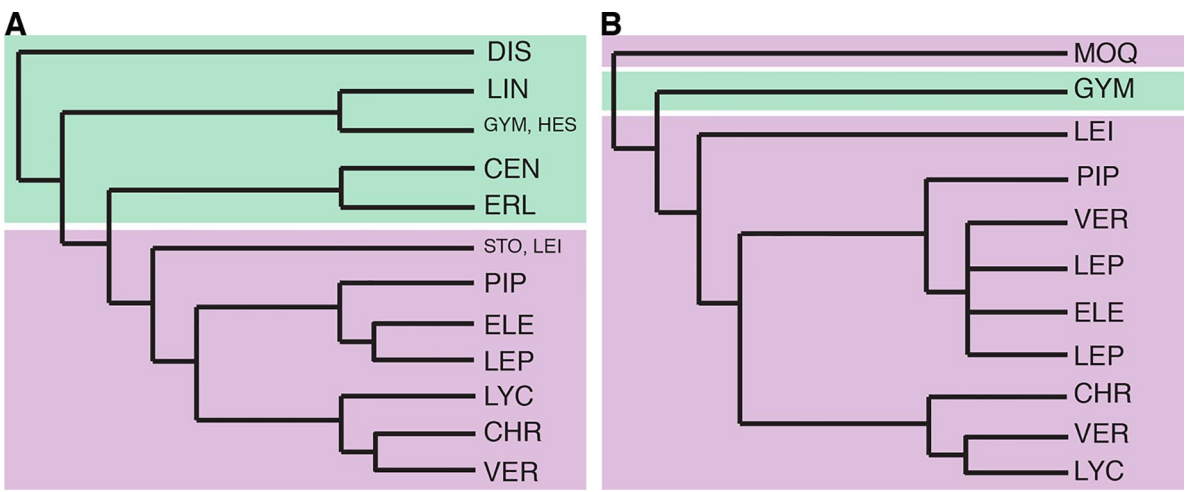

FIGURE 2 | Previous Vernonieae phylogenies. **(A)** redrawn from **Figure 2** in Keeley et al. (2007), Bayesian analysis of the combined molecular dataset. **(B)** redrawn from **Figure 2** in Loeuille et al. (2015a), strict consensus of 96 equally most parsimonious trees based on the combined molecular data. Branch lengths are illustrative, without real value. Green shading represents taxa with mostly Old World distribution and purple shading those with mostly New World distribution. CEN, Centrapalinae; CHR, Chrestinae; DIS, Distephaninae; ELE, Elephantopinae; ERL, Erlangeinae; GYM, Gymnantheminae; HES, Hesperomanniinae; MOQ, Moquinieae; LEI, Leiboldiinae; LEP, Lepidaploinae; LIN, Linziinae; LYC, Lychnophorinae; PIP, Piptocarphinae; STO, Stokesiinae; VER, Vernoniinae.

in genus *Lessingianthus*. The relative position of each clade also is different from that found in Keeley et al. (2007), especially with regard to the clades containing subtribes Lychnophorinae, Chrestinae, and *Vernonia* s.str. These relationships also vary depending on the dataset used, and the position of *Stokesia* also changes depending on the analysis, with it emerging with low support as a sister to subtribe Chrestinae, well within the New World clade and not in a transitional position, or as a sister to Leiboldiinae, as seen in Keeley et al. (2007).

Regarding the position of Vernonieae within Asteraceae, the tribe has usually been placed in Cichorioideae and is known to be closely related to Liabeae (Keeley et al., 2007; Panero and Funk, 2008). Relationships within Cichorioideae have always been unstable (Funk and Chan, 2009), with recent evidence that tribe Cichorieae might be more closely related to subfamily Asteroideae than to the rest of the tribes in Cichorioideae itself (Mandel et al., 2017; Mandel et al., 2019). In the megatree by Funk et al. (2009), the small South American tribe Moquinieae (**Figure 1B**) emerges in a polytomy with Vernonieae and *Distephanus* and also presents alternative placements in relation to both in an in-depth analysis of the relative positions of the tribes in Cichorieae (Funk and Chan, 2009), showing these relationships require further investigation.

Based on the hitherto known information about the phylogeny of Vernonieae and focusing on resolving some of the controversies between previous works, we carried out a phylogenetic study employing genomic methods, in order to: 1) understand the relationships among different subtribes in Vernonieae, especially among South American groups, 2) define the relationships among Moquinieae, *Distephanus*, and the core Vernonieae, and 3) understand the impact of different levels of missing data and of concatenated and pseudo-coalescence methods in the phylogenetic analysis.

MATERIAL AND METHODS

Outgroup Choice and Taxon Sampling

As Liabeae, Moquinieae, and *Distephanus* have been shown to be the sister groups to Vernonieae in previous works (Keeley and Robinson, 2009; Loeuille et al., 2015a), we chose as outgroup one taxon from Liabeae, *Munnozia gigantea*, and sampled as ingroup the only two representatives from tribe Moquinieae (*Moquinia racemosa* and *Pseudostifftia kingii*), *Distephanus ambonguensis*, and another 56 species representing 29 different genera from Vernonieae (4% of the species ascribed to the tribe). Taxa from 12 subtribes (from the 21 defined by Keeley and Robinson, 2009) were included, of which nine occur in South America and three are distributed in Africa/Asia. The sampling was focused on the three large South American clades that showed uncertain relationships based on previous studies (Keeley et al., 2007, Loeuille et al., 2015a). Sequences for 25 taxa were newly generated for this study, while sequences for the remaining 35 species were previously published elsewhere (Mandel et al., 2014; Mitchell et al., 2017; Mandel et al., 2019). A list of sampled species, herbarium vouchers, and publication status is presented in **Supplemental Material Table 1**.

DNA Extraction and Sequencing

Leaf samples were collected from live plants in the field and preserved in silica gel or extracted from herbarium sheets. Dried leaves were ground using a GenoGrinder 3000 (Spex® Sample Prep), and total DNA was extracted using E.Z.N.A.® SQ Plant DNA Kit from Omega Biotek, with addition of polyvinylpyrrolidone and ascorbic acid to the first extraction buffer (10-ml SQ1 buffer, 100-mg polyvinylpyrrolidone, 90-mg ascorbic acid). When necessary, the extracted DNA was cleaned with the E.Z.N.A.® Cycle Pure Kit from Omega Biotek to increase purity. Extracted samples were quantified using fluorometry (Qubit 3.0, ThermoFisher Scientific), diluted as necessary, and sheared to a target size of 400–500 bp using a sonicator (Covaris S series or QSonica Q500). DNA fragmentation was verified through electrophoresis in 1% agarose gels.

Libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs Inc.) with an initial concentration of at least 500 ng of total DNA, according to the manufacturer's instructions, using 15 cycles on the last amplification step. Final library concentrations and sizes were checked using Qubit and gel electrophoresis. Libraries were pooled in groups of four in equimolar concentration, containing 125 ng of each library, and target capture was performed using the MYbaits COS: Compositae/Asteraceae 1kv1 kit (Arbor Biosciences), using a 36-h incubation time and 15 cycles on the last amplification step. Details on the targets and method can be found in Mandel et al. (2014).

Quality checking with a Bioanalyzer instrument and sequencing were carried out at Macrogen Inc. (South Korea), in an Illumina HiSeq2500 device, in paired-end, high-throughput mode.

Sequence Assembly and Mapping

Trimming of Illumina adaptors was carried out using Trimmomatic (Bolger et al., 2014), and reads were assembled into contigs using SPAdes (Bankevich et al., 2012), with kmer lengths of 21, 33, 55, 77, 99, and 127. The sequences were matched back to the original probes using the phyluce pipeline (Faircloth, 2016), which generated individual alignments for each one of the original targeted regions. These alignments were then concatenated to generate two different matrices for phylogenetic analysis, using the "phyluce_align_get_only_loci_with_min_taxa" script within the phyluce pipeline, specifying different degrees of completeness in relation to number of loci contained in the final matrix. One matrix contains all loci recovered for all taxa (herewith called total matrix), and the other contains only loci that were recovered for at least 75% of the taxa (called 75% matrix). This approach was chosen to study the effect that different levels of missing data would have over tree topology and statistical support. General information about the matrices was obtained using AMAS (Borowiec, 2016) and files generated by the phyluce pipeline.

Phylogenetic Analysis

All analyses described were carried out with both datasets, total and 75%, containing invariable characters, using *M. gigantea* as

outgroup. The resulting trees are referred to as "total tree" and "75% tree" throughout the results and discussion. Molecular evolution models were evaluated in jModelTest2 (Guindon and Gascuel, 2003; Darriba et al., 2012), using the corrected Akaike Information Criterion and Bayesian Information Criterion to choose between models. The chosen model was GTR + I + G for both matrices and both information criteria. Maximum likelihood (ML) analyses were run on RAxML (Stamatakis, 2014) in the rapid bootstrapping mode, always using 1,000 bootstraps and 25 threads.

The multispecies pseudocoalescence model was evaluated in ASTRAL III (Zhang et al., 2018), using unrooted gene trees generated from the individual locus matrices. Individual evolution models for each gene matrix were obtained with PartitionFinder v.1.1.0, in the RAxML version with rcluster search option and Akaike Information Criterion, with unlinked branch lengths (Stamatakis, 2006; Lanfear et al., 2012; Lanfear et al., 2014). Gene trees were obtained in RAxML, with 100 bootstraps for each matrix. Two different species trees were obtained from the gene trees: one using all recovered loci and other using only loci that were recovered for 75% of the taxa. Branch support was calculated using local posterior probabilities (LPP).

The presence of gene tree conflict and concordance in the pseudocoalescence analyses was checked using PhyParts (Smith et al., 2015). Gene trees used as input for ASTRAL and the resulting species tree generated by the program were unrooted and, thus, had to be rooted to be used as input in PhyParts, which was done using the program pxrr in the package phyx (Brown et al., 2017). Species trees were rooted having *M. gigantea* (from Liabeae) as outgroup. Because the incomplete recovery of loci across taxa leads to several missing taxa in each gene tree, a hierarchical strategy was used to root the gene trees, selecting the outgroup in the following order: *M. gigantea*, *Distephanus ambongensis*, *P. kingii*, *M. racemosa, Vernoniastrum ambiguum*, *Baccharoides anthelmintica*, *Gymnanthemum amygdalinum*, *Centrapalus pauciflorus*, and *Stokesia laevis.* The results from PhyParts were used as input in the phypartspiecharts.py script (Johnson, 2017), to generate a species tree with pie charts in each node showing the proportion of concordant gene trees and conflicting topologies.

The occurrence of long-branch attraction (LBA) was tested using TreeShrink (Mai and Mirarab, 2018), both in the species trees generated by maximum likelihood and pseudocoalescence analyses and in the gene trees used as input to ASTRAL, using a false-positive error rate (α) of 0.05. Pseudocoalescence analyses were rerun with the treated gene trees to account for possible changes in topology and support values.

Topological Comparison

The topologies obtained with the different analyses and datasets were compared using the adjusted Robinson Foulds distance, as outline in Mitchell et al. (2017) and Herrando-Moraira and The Cardueae Radiations Group (2018). Robinson Foulds distances were calculated in PAUP* v4.0a (Swofford, 2003) for all pairwise comparisons of the six topologies (the total and 75% dataset for each of three analyses: ML, pseudocoalescence, and pseudocoalescence with gene trees treated with TreeShrink analyses) and then manually adjusted using RFadj = RF/(2n - 6), where *n* is the number of nodes in the tree. RFadj ranges from 0 (same topology) to 1 (completely discordant topology). The multidimensional scaling approach implemented in R was used to visualize all the trees in the same treespace, based on the RFadj values, using the function "cmdscale" in the package "stats."

RESULTS

Overview and General Trends

The sequencing generated approximately 902 million reads and approximately 89 billion nucleotides (4 million to 33 million reads per sample). The total matrix contains 61 taxa and has an extension of 729,969 characters, including 707 of the markers contained in the probe set, with 74.9% missing data. The 75% matrix has 61 taxa as well, but the matrix length is of 113,347 characters, containing 89 loci and 34.9% missing data. The number of loci recovered for each taxon varied from 79 in *M. racemosa* to 492 in *C. pauciflorus*, with a median of 249 loci. Although there is a drastic reduction in the number of variable and parsimony-informative sites in the 75% matrix compared with the total matrix, proportionally, the 75% matrix has more parsimony-informative sites (19% against 13%). Comprehensive data for the recovered loci and alignments are found in **Table 1** and **Supplemental Material Table 2**. Raw data are deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive, under BioProjects PRJNA540287 and PRJNA546287.

Overall, the four analyses are remarkably consistent, presenting similar topologies and high statistical support (**Figures 2**, **3** and **Supplemental Material Figures 1–3**). Some of the general trends found in all analyses are the position of *Distephanus* and Moquinieae in relation to Vernonieae, these three species form a clade with *Distephanus* as sister group to the other two, although with low support in the ML analysis and high support in the pseudocoalescence (support for total/75% trees: ML bootstrap:

TABLE 1 | Comparison of the composition of the total and 75% datasets.

	Total matrix	75% matrix
Number of taxa	61	61
Number of recovered loci	707	89
Length of the concatenated matrix	729,969 bp	113,347 bp
Number of variable sites	235,126 bp (32%)	40,784 bp (35%)
Number of parsimony informative sites	101,707 bp (13%)	22,650 bp (19%)
Proportion of missing data in the concatenated matrix	74.6%	35.7%
Proportion of identical sites in the concatenated matrix	38.5%	22.1%
Average of recovered loci per species (sd; min–max)	247 (87; 79–492)	72 (14; 25–88)
Average number of species recovered per loci (sd; min–max)	20 (15; 3–56)	49 (2; 45–56)
Average sequence length	1,032 (770; 167–10,337)	1,273 (1,117; 372–9,205)

bp, base pairs; sd, standard deviation; min, minimum; max, maximum.

5%, 92%, LPP: 1, 1). Three of the sampled African species form a consistent clade, recovered in all analyses, with *G. amygdalinum* as the sister taxon to *V. ambiguum* and *B. anthelmintica,* with maximum statistical support in all cases. Also, there is an inconsistency in the position of *C. pauciflorus* and *S. laevis* as sister to the South American clade, probably due to the incomplete sampling of African taxa and Mexican subtribe Leiboldiinae. Subtribe Chrestinae, composed only by *Chresta* (**Figure 1L**), is consistently monophyletic with high statistical support (support for total/75% trees: ML bootstrap: 100%, 100%, local PP: 1, 1), and its sister group is a clade formed by *Heterocypsela andersonii* (**Figure 1K**) + *Vernonia s.str.* + *Vernonanthura*, also with high statistical support (ML bootstrap: 100%, 100%, local PP: 1, 0.99). The relative position of *Chresta exsucca*, *C. scapigera*, and *C. sphaerocephala* varies in the analysis depending on the dataset.

Subtribe Vernoniinae was recovered as non-monophyletic in all analyses; instead, they are split in two clades: *Vernonia* and *Vernonanthura* grouped as the sister clade to Chrestinae and *Cyrtocymura* as sister taxon to Lepidaploinae (**Figures 3** and **4**). Subtribe Lepidaploinae also emerges as non-monophyletic, and although all the species are grouped into a large clade, *Stilpnopappus* and *Strophopappus* (**Figure 1I**) form the sister clade of the Elephantopinae, and *Lepidaploa* and *Lessingianthus* (**Figure 1H**) are in a different clade that also contains *Cyrtocymura*. The monotypic genus *Soaresia* (**Figure 1J**) is included in the Elephantopinae (**Figure 3**).

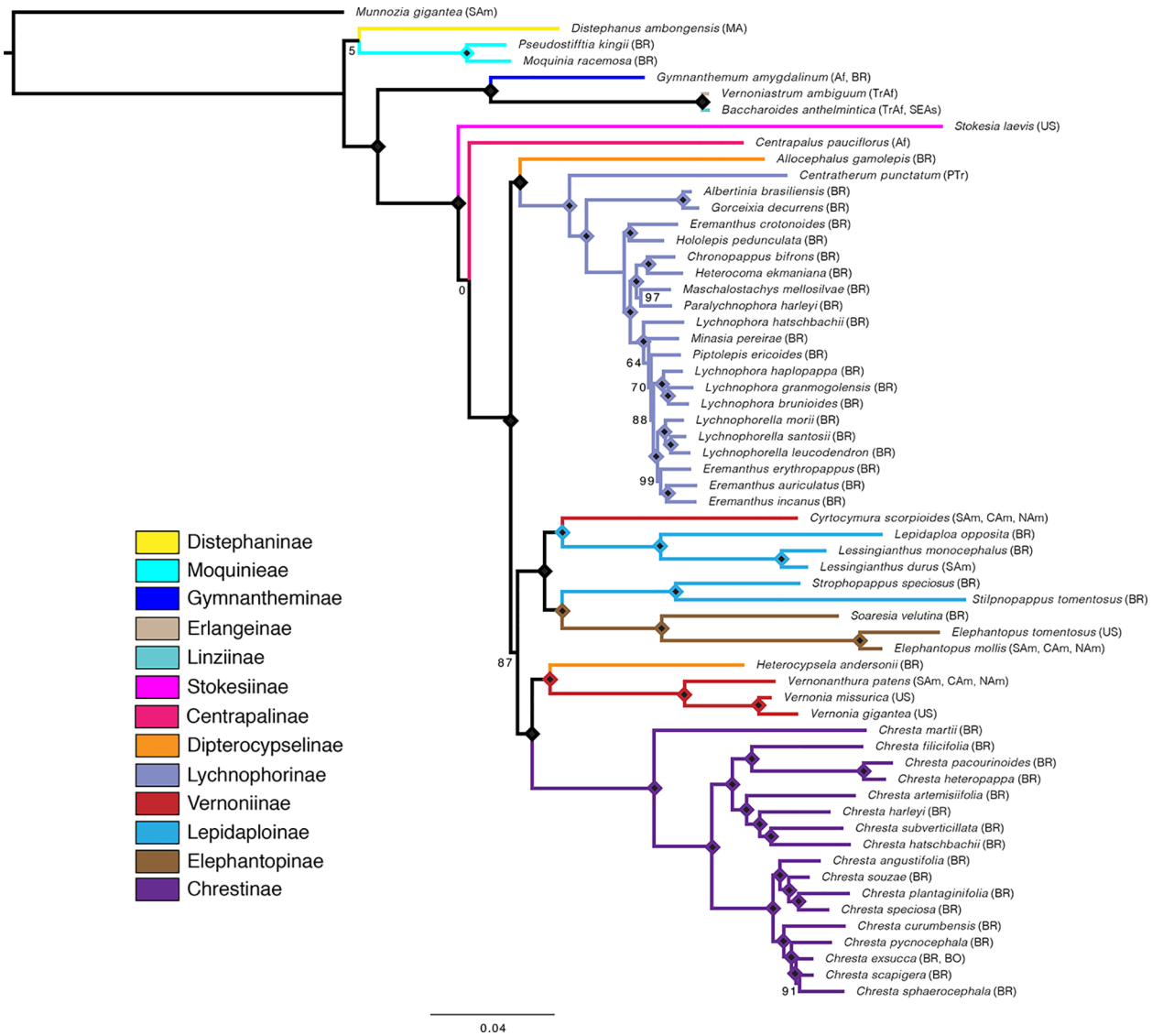

FIGURE 3 | Maximum likelihood tree obtained with the total matrix (707 loci, 729,969 characters), with model GTR + G + I, with 1,000 bootstrap replicates. Diamonds indicate bootstrap value of 100%. Subtribes are coded by color. Geographical distribution indicated in parenthesis: Af, Africa; BO, Bolivia; BR, Brazil; CAm, Central America; MA, Madagascar; NAm, North America; PTr, Pantropical; SAm, South America; SEAs, Southeast Asia; TrAf, Tropical Africa; US, United States of America. Distribution data obtained from Keeley and Robinson (2009), Robinson (1999a, 1999b).

The two species belonging to subtribe Dipterocypselinae emerge in two distantly related clades, *Heterocypsela* is in a clade with part of the Vernoniinae, and *Allocephalus gamolepis* emerges as sister group of the Lychnophorinae, rendering Dipterocypselinae non-monophyletic.

Lychnophorinae is recovered as monophyletic in all analyses, with some of the relationships within this subtribe being stable, such as the clades formed by *Albertinia brasiliensis* and *Gorceixia decurrens* and *Eremanthus crotonoides* and *Hololepis pedunculata* (**Figure 1F**). There is also a clade formed by four species divided into two clades: *Paralychnophora harleyi* + *Maschalostachys mellosilvae* and *Chronopappus bifrons* + *Heterocoma ekmaniana*. The position of *Minasia* and *Piptolepis* in relation to *Lychnophora* and *Eremanthus* varies between the 75% and total datasets, in both the ML and pseudocoalescence analysis.

Maximum Likelihood

The main difference between the trees obtained in the ML analysis is the relative position of the three large South American clades. In the total ML tree (**Figure 3**), the clade formed by Elephantopinae + Lepidaploinae + Vernoniinae (ELE + LEP + VER) is the sister clade to Chrestinae + Vernoniinae (CHR + VER), and both together form the sister clade of Lychnophorinae (LYC). In the 75% tree (**Supplemental Material Figure 1**), LYC emerges the sister group of CHR + VER, and ELE + LEP + VER is the sister group of the remainder. However, in the 75% tree, these relationships all have total support, while in the total tree, the CHR + VER and ELE + LEP + VER node has 87% of bootstrap support.

The overall support is higher in the total tree, with nine nodes showing support below 100%, while the 75% tree has 13 nodes with lower support. Some nodes with lower support are shared by both trees, such as the basal node in the *Distephanus* + Moquinieae clade, which has lower support in the total tree. Also, the position of *Stokesia* and *Centrapalus* changes in both trees. In the total tree, *Stokesia* emerges before *Centrapalus*, and the node between *Centrapalus* and the South American clade presents no support. In the 75% tree, they are inverted, and the node between *Stokesia* and the South American clade has 43% support.

The number of nodes with lower support within Lychnophorinae and Chrestinae increases in the 75% tree, and there are also changes in topology within these clades, especially in the innermost clade of Lychnophorinae. The analyses with TreeShrink in both trees indicate a possible LBA case with *M. gigantea*, which is the outgroup. Rerunning the analysis with the same α level and removing the outgroup indicate a possible case of LBA with *S. laevis* in the 75% tree, which might explain the inverted position of this taxa and *C. pauciflorus* between the two trees.

Multispecies Pseudocoalescence

The pseudocoalescence analysis with all loci included 645 gene trees, while the analysis containing only loci that where recovered for at least 75% of the taxa contained 87 gene trees. The normalized quartet score for both datasets was 0.84. Overall, LPP values were strongly affected by reducing the number of loci in the analysis, and the total tree has 12 nodes with support below 1, while the 75% tree has 25 nodes with support below 1.

Differently from the maximum likelihood analyses, there is no variation in the backbone topology between both analyses, with the trees presenting the same relationship among the three large South American clades, where CHR + VER and ELE + LEP + VER are sister clades and LYC is the sister group of this larger clade, in accordance to the topology in the ML total tree. However, the support in the CHR + VER and ELE + LEP + VER node was low in both trees (LPP total/75% tree: 0.85/0.45). There is variation in the topology within clades, especially within Lychnophorinae and in one clade in Chrestinae. There is no variation in the position of *Centrapalus* and *Stokesia*, with *Centrapalus* emerging before *Stokesia*, with high support in both cases (LPP: 1/1, 0.99/0.95) (**Figures 4B**, **C**, **Supplemental Material Figures 2 and 3**).

Removing taxa that could potentially cause LBA from the gene trees with TreeShrink did not change the topology of the resulting species trees and had confounding effects on overall support. In the total tree (**Supplemental Material Figure 4**), half of the taxa were removed from at least 10 gene trees each, and *B. anthelmintica* and *C. pauciflorus* were removed from 21 and 24 gene trees, respectively (**Supplemental Material Table 1**). The number of nodes with LPP < 1 remained the same (12), but in some of these nodes, the support decreased, such as the node containing CHR + VER and ELE + LEP + VER, in which the support fell from 0.85 to 0.73. In the loci contained in the 75% tree (**Supplemental Material Figure 5**), *M. gigantea* was removed from 10 gene trees and *S. laevis* from 5 gene trees; six other taxa were removed from one tree each (**Supplemental Material Table 3**). The number of nodes with LPP < 1 also remained the same, and the biggest change in support occurred in the *Distaphanus* + Moquinieae node, which fell from 1 to 0.85.

Even though the statistical support generally fell in the 75% tree, the gene tree concordance analysis shows there is less discordance between gene trees than in the total tree. In the total tree, 91% of the nodes show that more than 50% of the gene trees are non-informative for that node, and only small proportions of the trees are concordant (**Figure 5**). The backbone of the tree has lower proportions of non-informative gene trees and also shows concordance with alternative topologies. Nodes within Lychnophorinae are overall more uninformative than in other parts of the tree, also corresponding to the region where support is lower. The 75% tree shows smaller proportions of non-informative gene trees for each node, and 36% of the nodes show a proportion of 50% or more of concordant gene trees (**Figure 6**). The backbone shows higher proportions of concordance, and most of the nodes that showed higher proportions of uninformativeness in the total tree show concordance with alternative topologies in the 75% tree.

Topological Comparison

Discordant tree topologies were recovered, especially when comparing the two different datasets, including a significant change in the backbone between the two ML topologies. The RFadj values were generally low (**Table 2**), with the largest difference being between the two ML analyses (RFadj = 0.13).

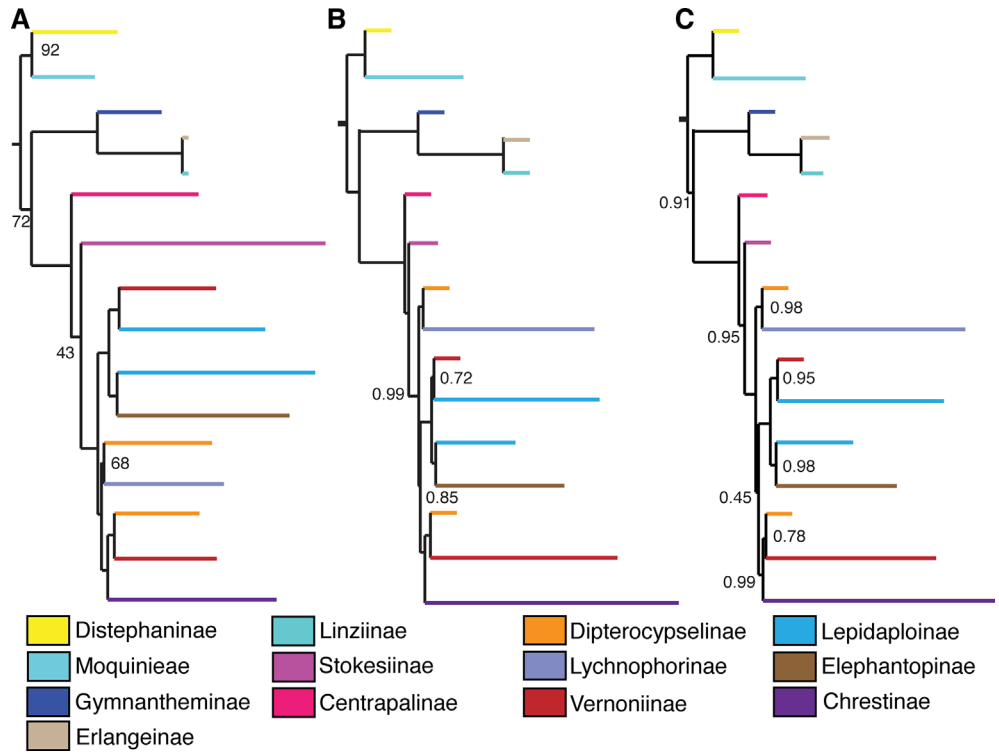

FIGURE 4 | Comparison between backbone trees obtained with different analysis methods and datasets. **(A)** 75% matrix, maximum likelihood. **(B)** Total matrix, pseudocoalescence. **(C)** 75% matrix, pseudocoalescence. Nodes without a value have 100% bootstrap support or local posterior probability of 1. Subtribes are coded by color.

The smallest difference was between the total ASTRAL analysis with the total ASTRAL analysis with TreeShrink, which were completely concordant (RFadj = 0.0). Overall the comparisons between the two different datasets had higher discordance, possibly indicating that the dataset, not the analysis method, was driving the differences in topologies, as seen in **Figure 7**. Using TreeShrink to remove possible anomalous taxa that could cause LBA before running pseudocoalescence analysis did not cause drastic differences in "before and after" topologies.

DISCUSSION

Agreement Between Datasets and Analysis

The results obtained with different analyses were overall consistent, and incongruences seem to be more related to the dataset used than to the type of analysis, as indicated by RFadj values. As the level of missing data is a frequent problem in studies based on multiple markers (Huang and Lacey Knowles, 2016), we used a 75% matrix as a strategy to try to understand the effect that the high level of missing data might have on the topology and support. The effect of missing data in phylogenetic analyses has been addressed at least since fossils were included in them (Donoghue et al., 1989), and missing data are being increasingly discussed as larger datasets continue to appear. One view on the problem is that missing data do not influence the outcomes so strongly when a sufficient amount of characters has been sampled (Wiens, 2003; Wiens and Morrill, 2011).

In our analyses, reducing the number of markers decreased overall support on the trees, especially on the coalescence tree. The 75% ML analysis, besides presenting lower support, presents a major change in the position of the large South American clades, with additional changes within the clades. In the coalescence analysis, the position of the major clades remains the same, although the number of nodes with low support doubles in the 75% tree. This finding may indicate that in these two analyses, the full dataset helps to resolve internal nodes and gives more characters that support the relationships established by the cleaner dataset. However, the results of the partition analysis with PhyParts showed that removing the gene trees that are more incomplete in terms of represented taxa did improve the agreement between gene trees and species trees.

Other explanations for the differences found in the internal relationships are that some clades include a large variation in the number of loci recovered or variation in what loci were recovered in each taxon, low variation among taxa in the recovered loci, and also inadequate sampling. In Chrestinae, the most likely reason is that the recovered loci are too similar among the three species whose positions vary in different analyses, as the genus was well sampled (17 of 18 species) and the number of loci recovered for each taxon was fairly similar

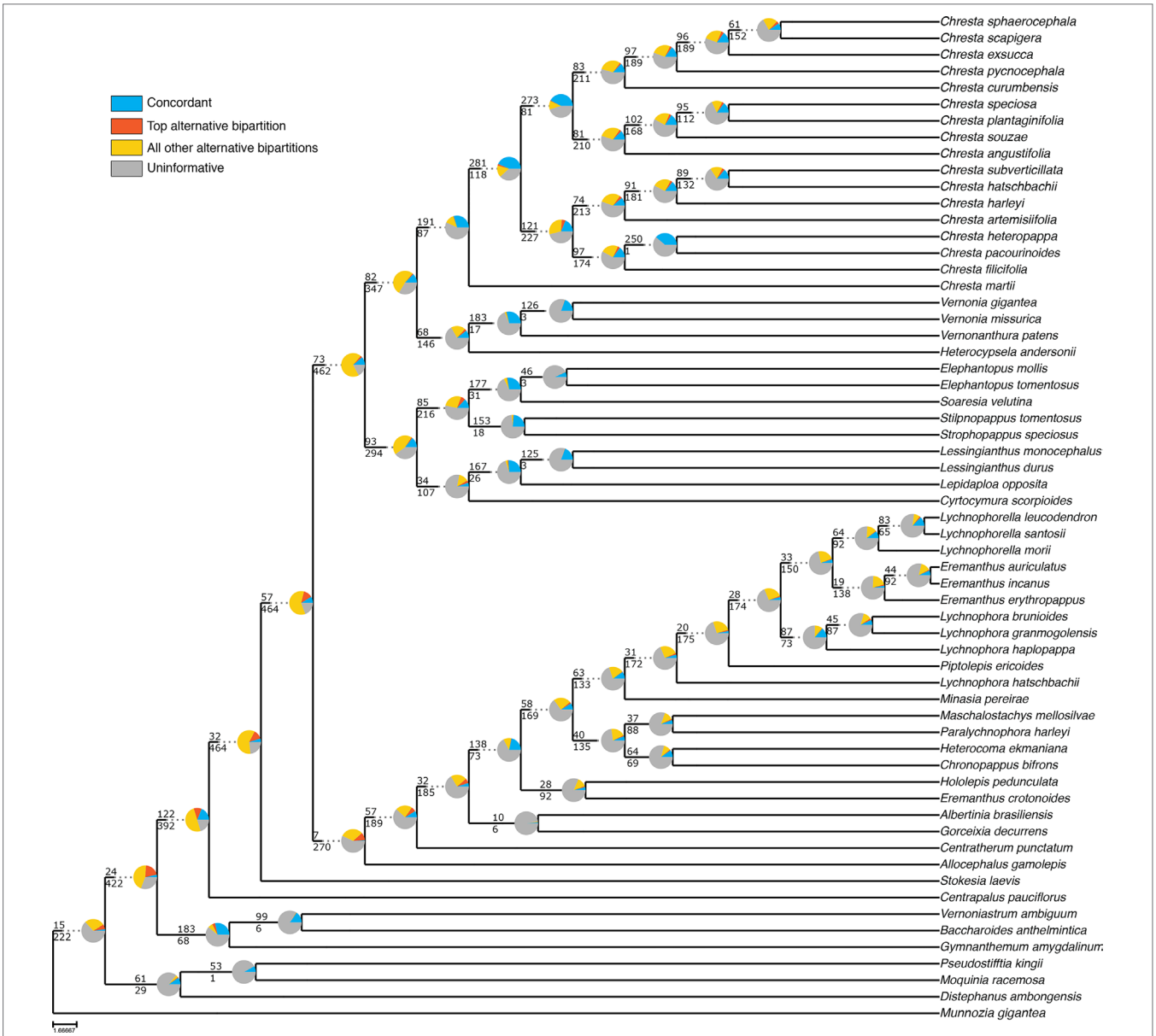

FIGURE 5 | ASTRAL analysis topology of the total dataset showing a summary of concordant and discordant gene trees. For each branch, the top number indicates the number of concordant gene trees and the bottom number the number of conflicting gene trees. The pie charts indicate the proportion of gene trees that support that clade (blue), the proportion that supports the main alternative for that clade (orange), the proportion that supports all other topologies (yellow) or the proportion of uninformative gene trees for that clade (gray).

(271 to 307 loci with 203 being recovered for all three species). In Lychnophorinae, changes in internal relationships are likely due to the poor sampling of this diverse subtribe (only ~18% of species sampled), and also, the fact that this subtribe seems to have diversified in a short time frame, possibly leading to low sequence divergence.

As new methods for obtaining large numbers of loci have appeared, the discussion about appropriate methods for phylogenetic inference has become a debated topic, with multiple authors advocating for the multispecies coalescence method as a more precise and biologically correct approach, as it incorporates gene tree heterogeneity that usually is ignored in analysis of concatenated matrices (Edwards et al., 2016). Overall, the phylogenetic relationships reported here are in agreement, including those recovered with different analytical methods. However, partition analysis indicates strong disagreement among gene trees and an abundance of uninformative gene trees, which improved with removal of loci that were recovered for less than 75% of the taxa present in tree.

As previously shown in a study in Cardueae, another tribe in Asteraceae, the pseudocoalescence method tends to produce trees that are more congruent in their topologies when different

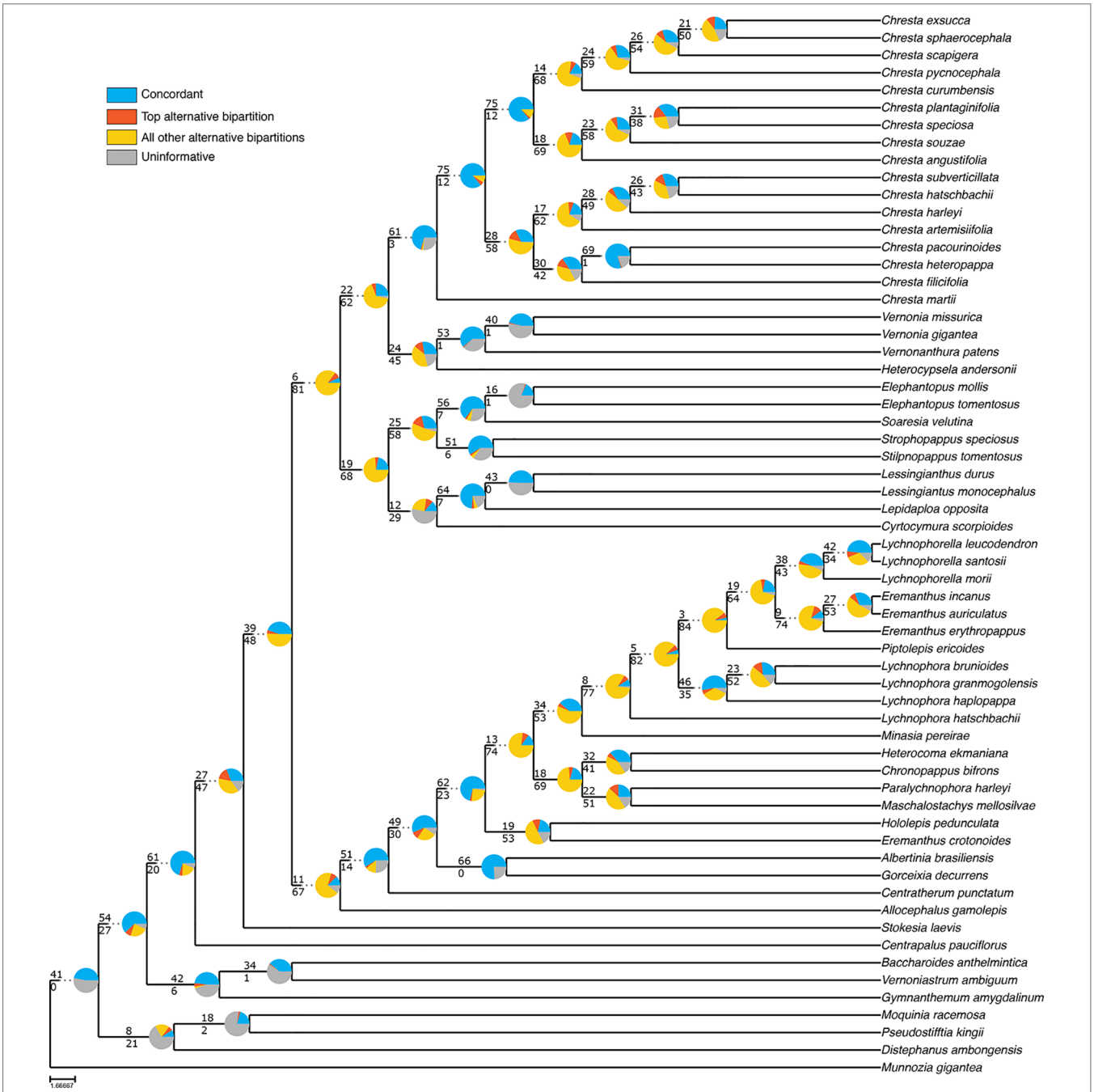

FIGURE 6 | ASTRAL analysis topology of the 75% dataset showing a summary of concordant and discordant gene trees. For each branch, the top number indicates the number of concordant gene trees and the bottom number the number of conflicting gene trees. The pie charts indicate the proportion of gene trees that support that clade (blue), the proportion that supports the main alternative for that clade (orange), the proportion that supports all other topologies (yellow) or the proportion of uninformative gene trees for that clade (gray).

datasets are used (Herrando-Moraira and The Cardueae Radiations Group, 2018). We found a similar result, with the two coalescence analyses presenting only small changes in topology in internal nodes, while the reduced dataset in the concatenation analysis produced a tree with a significant change in the backbone topology. Overall support in coalescence trees seems to be largely improved by keeping a higher number of loci, even if it increases the percentage of missing data (Liu et al., 2015, Herrando-Moraira and The Cardueae Radiations Group, 2018), a result that we also observe in the current study. Removing taxa that could potentially cause LBA does not improve support in either of our trees. These results are in agreement with simulation studies, which found that pseudocoalescence methods based on gene tree topology, such as ASTRAL, are resilient to LBA effects (Roch et al., 2019).

TABLE 2 | Pairwise adjusted Robinson Foulds distance between each pair of tree topologies.

Tree topology	ML total	ML 75%	ASTRAL total	ASTRAL 75%	TreeShrink total	TreeShrink 75%
ML total	0	–	–	–	–	–
ML 75%	0.13	0	–	–	–	–
ASTRAL total	0.04	0.09	0	–	–	–
ASTRAL 75%	0.07	0.07	0.04	0	–	–
TreeShrink total	0.04	0.09	0	0.04	0	–
TreeShrink 75%	0.05	0.09	0.02	0.02	0.02	0

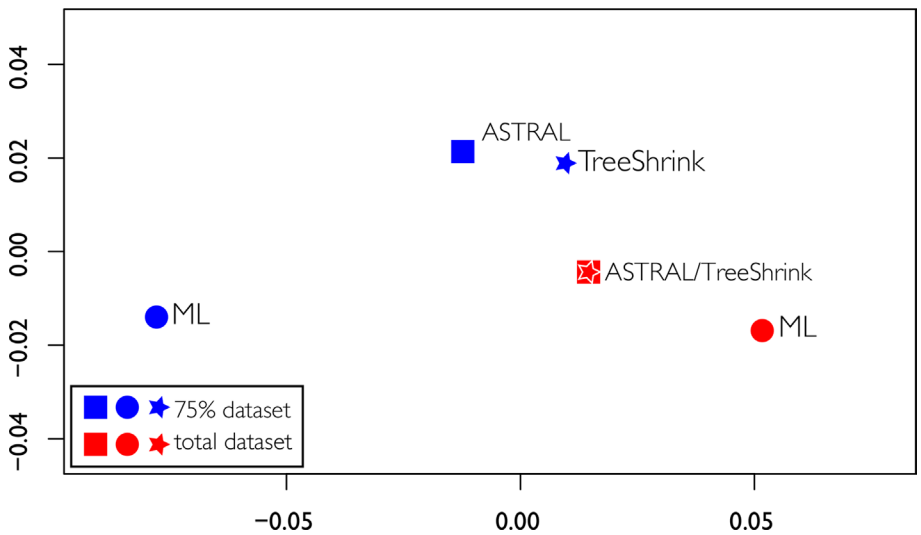

FIGURE 7 | Tree space of the multidimensional scaling of pairwise comparisons of Robinson Foulds distances among the four topologies obtained. ML trees are represented as circles, ASTRAL trees as squares, and ASTRAL trees obtained with gene trees treated with TreeShrink as stars.

The presence of paralogs in Asteraceae is abundant and the family has an extensively studied history of whole genome duplications (Barker et al., 2008; Barker et al., 2016; Huang et al., 2016). The probes developed by Mandel et al. (2014) used here contain a set of mostly orthologous genes; however, the phyluce pipeline still points out the recovery of possible paralogous loci in varying degrees across the species. As the probes were originally based on taxa distant from Vernonieae, we opted for completely removing any loci that could possibly present paralogy, as orthology assessment would likely be impaired by phylogenetic distance.

Relationships Among Moquinieae, *Distephanus*, and Core Vernonieae

The present work is the first one focused on the Vernonieae that included both Moquinieae and *Distephanus*. Keeley et al. (2007) used *Distephanus* as an outgroup, while Loeuille et al. (2015a) included *Moquinia* as an outgroup. Funk and Chan (2009) investigated the influence of including different tribes and using different outgroups in the relationships within Cichorieae and usually recovered Moquinieae as the sister to the core Vernonieae, while *Distephanus* usually emerges as sister taxon to Moquinieae plus Vernonieae. Here, we present a different relationship, consistently recovered in all our trees, where *Distephanus* and Moquinieae form a clade that is sister group of all Vernonieae. Curiously, in a recently published phylogeny for the family, where part of the data presented here is also included, *Distephanus,* Moquinieae, and Vernonieae emerge sequentially in all analyses (Mandel et al., 2019). It is possible that the sampling of only one representative (*M. gigantea*) from the 165 species of Liabeae (Dillon et al., 2009) as an outgroup biased the analysis and artificially created this clade containing Moquinieae and *Distephanus*.

The two members of Moquinieae, composing two monotypic genera, have an extensive taxonomic history, due to their unusual morphology. Although they present many similarities with the Vernonieae, especially in the homogamous heads and purple florets, the inflorescence, style, and pollen morphology are starkly different from those usually found in Vernonieae and other cichorioid tribes. *M. racemosa* was firstly placed with the Gochnatieae, while *P. kingii* was initially described in Vernonieae. The two species were synonymized into *Moquinia* and placed in Vernonieae in the 1990s (Gamerro, 1990), and Robinson (1994) later placed them as separate genera in their own tribe Moquinieae. *Distephanus* also presents an unusual morphology,

despite being recognized as part of the Vernonieae, with yellow flowers and trinervate leaf venation, being first placed in subtribe Liabinae (now tribe Liabeae) in the Senecioneae (Keeley and Robinson, 2009).

The phylogenetic position of this species-poor clade (tribe Moquinieae and *Distephanus*) leading to the species-rich Vernonieae potentially indicates an interesting and complicated biogeographic history, likely with multiple events of colonization of Africa and South America and extinction of lineages, as the outgroup Liabeae has an Andean distribution, while Moquinieae is exclusively Brazilian and the 50 species of *Distephanus* are distributed in Africa, India, and southern Asia. The African genera of Vernonieae have consistently been recovered as a grade leading to the New World clade (Keeley and Turner, 1990; Keeley et al., 2007), possibly indicating an initial diversification of the tribe in Africa before moving to South America again, which is in agreement with recent work (Mandel et al., 2019). Nevertheless, a detailed biogeographic study of the tribe and its closest relatives is still lacking.

Relationships Within Vernonieae and Agreement With Past Phylogenies

Relationships within Vernonieae, especially within the South American clades, were partially contradictory in previous phylogenies (Keeley et al., 2007; Loeuille et al., 2015a) and, even after the present study, are still not completely understood. Keeley's work (2007) has the most complete sampling in terms of genera and geographic distribution, especially regarding African and Asian genera, while the phylogeny by Loeuille et al. (2015a) expands the sampling of South American groups. Overall, the trees presented here are more similar to those found in Keeley et al. (2007).

The position of the monotypic *Stokesia* in the tribe is still a point of contention. In Keeley et al. (2007) and in the Bayesian analysis in Loeuille et al. (2015a), it is in the transition from the African to the South American Vernonieae as in our study, although in a clade with Mexican and Asian taxa. The anomalous morphology of the florets in this species, which are ligulate, and its isolated distribution in Southeastern USA might indicate that it is a leftover from a lineage that went through massive extinction, a pattern that seems frequent in Vernonieae with its abundance of monotypic genera (Keeley and Robinson, 2009).

Regarding the relationships in the South American clade, although the backbone presents wide variation among different analyses, some internal relationships remain stable. Both Keeley et al. (2007) and Loeuille et al. (2015a) recovered the same relationship between Elephantopinae and part of the Lepidaploinae. Although *Elephantopus* presents pantropical distribution, it is nested within the South American clade, with both our present work and Loeuille et al. (2015a) recovering the monotypic and strictly Brazilian *Soaresia* as its sister taxon, indicating a possible late migration from South America to other continents. Loeuille et al. (2015a) also showed the presence of Vernoniinae members, specifically *Cyrtocymura,* intermingled in Lepidaploinae, similar to the topology that we recovered here.

Keeley et al. (2007) showed a clade formed by Chrestinae and part of the Vernoniinae (*Vernonia* and *Vernonanthura*), as well as their relationship with *Heterocypsela*. In our work, we recovered a clade formed by *Heterocypsela, Vernonia*, and *Vernonanthura* as sister to Chrestinae, while in Keeley et al. (2007), *Heterocypsela* emerges as most closely related to *Chresta*. This previous work included two genera not sampled here, *Tephrothamnus* and *Eirmocephala*, from South and Central America, which could change the relationships we found if included. The relationship of *Chresta* with other Vernonieae has always been unclear (Robinson, 1992), as the genus presents secondary heads, which approximate it to the Lychnophorinae but also pollen and anther appendage features (Robinson, 1999a) that suggest a closer relationship to Vernoniinae.

Loeuille et al. (2015a) postulated the multiple origins of syncephaly in the Vernonieae, deeming classifications based on this character artificial. In the trees presented here, *Chresta* indeed is closer to other taxa lacking secondary heads than to Lychnophorinae, indicating the complex evolution of secondary heads, possibly through different developmental steps.

The relationship of the large clade formed by CHR + VER with the other Vernonieae varies depending on the analysis and dataset, although most trees agree with CHR + VER being the sister group of ELE + LEP + VER (**Figure 4**), although with low support. The exception is the ML analysis with the 75% dataset (**Figure 4A**), which shows CHR + VER as the sister group of Lychnophorinae. In all other analyses, LYC emerges as the sister group of (CHR + VER) + (ELE + LEP + VER). Keeley's phylogeny (2007) agrees with our 75% ML analysis, with (CHR + VER) + LYC and ELE + LEP + VER as the sister clade of this larger clade.

In Loeuille's work (2015a), Chrestinae and *Stokesia* emerge as sister to a clade formed by LYC and *Vernonia* + *Vernonanthura*. The bulk of LEP groups with ELE and some other VER, forming the sister clade of CHR + (*Vernonia* + *Vernonanthura* + LYC). None of the trees in the present work support these relationships.

When subtribe Chrestinae was created (Robinson, 1992; Robinson, 1999a), the monotypic genus *Soaresia* from Central Brazil was placed in it due to some morphological similarities, mainly the presence of secondary head and pollen type. However, in Loeuille's work (2015a), *Soaresia* emerges as the sister taxon of the Elephantopinae, with the same relation shown in all analyses presented here. In fact, *Soaresia* has morphological affinities to *Elephantopus,* such as the bristle-like awls that compose the pappus and the unbranched trichomes, further supporting its transference to subtribe Elephantopinae (Loeuille et al., 2015a).

Also, the analyses presented here do not support the monophyly of subtribe Dipterocypselinae. This subtribe was created to accommodate two monotypic genera that present dimorphic cypsela (*Dipterocypsela* and *Heterocypsela*) and a third monotypic genus (*Manyonia*) without dimorphic cypsela (Keeley and Robinson, 2009), with a fourth monotypic genus (*Allocephalus*) with dimorphic cypsela being added later (Bringel Jr et al., 2011). We sampled only the two Brazilian

representatives of the subtribe, *Heterocypsela* and *Allocephalus*, both from Central Brazil and growing on limestone outcrops. *Dipterocypsela* is found on Northern Colombia, also on limestone outcrops (Blake, 1945). *Manyonia* does not present fruit dimorphism, but the inflorescence structure and the pattern of the cells on the cypsela walls placed it close to *Heterocypsela* and *Dipterocypsela* (Robinson, 1999b), regardless of this species being known only in Tanzania. *Heterocypsela* and *Allocephalus* fall in distant places in our trees, in the Vernoniinae and Lychnophorinae, respectively. Due to the morphological singularities of these four genera, their placement within Vernonieae subtribes has always been putative at best (Blake, 1945, Robinson, 1999b), and its status as a subtribe should be reevaluated, depending on the inclusion of *Dipterocypsela* and *Manyonia* in future analyses.

Another finding from our analyses is the non-monophyly of both Vernoniinae and Lepidaploinae. As sampled here, Lepidaploinae terminals emerged in two clades, one including *Cyrtocymura*, which is currently placed in Vernoniinae, and another sister to Elephantopinae. Vernoniinae terminals also emerged separated, with *Vernonia* and *Vernonanthura* being sister to *Chresta*, and *Cyrtocymura* grouping with the LEP + ELE. These separations had already been shown in Loeuille's analysis (2015a), although with lower resolution and support. Lepidaploinae was initially included as a complex of genera within Vernoniinae (Robinson, 1999a), later being separated due to complex combinations of micro- and macrocharacters (Keeley and Robinson, 2009), such as the echinolophate pollen and the seriate-cymose inflorescences. Although combinations of characters can be useful for identification of genera and species, it is becoming clear that many of them are homoplastic, producing classifications that do not reflect the evolutionary history, and this seems to be the case in the infra-tribal classification in Vernonieae, which will have to be reevaluated as more inclusive analyses become available.

Regarding Lychnophorinae, the relations uncovered here slightly differ from those seen in Loeuille et al. (2015b); however, these differences are difficult to interpret due to our low taxonomic sampling, which includes only a few representatives from each major clade within it. As previously shown by Loeuille et al. (2015a, 2015b), Centratherinae emerges as the sister taxon of all other Lychnophorinae and is now considered a synonym (Loeuille et al., 2019), as well as Sipolisiinae, whose members emerge in several positions within Lychnophorinae. The monotypic *Allocephalus*, not included in previous phylogenies, here emerges as sister to the rest of Lychnophorinae. It displays various plesiomorphic features of Lychnophorinae: herbaceous habit (*Centratherum*), T-shaped trichomes (*Albertinia*, *Centratherum,* etc.), and heads in dense glomerules (*Blanchetia*, *Gorceixia*). It shares with *Albertinia* a style with basal node (feature uncommon in Lychnophorinae) and especially, as noted by Bringel Jr et al. (2011), an involucre with fused phyllaries.

This peculiar involucre sheds an interesting light on the origin of the unique alveolate receptacle of *Albertinia* that has been variously interpreted: Candolle (1836) assumed that *Albertinia* had one floret per capitulum and fused capitula as in *Eremanthus* and *Lychnophora,* but since Schultz-Bipontinus (1861, 1863), *Albertinia* capitula are interpreted as multiflowered and the receptacle surface with deep holes (alveolae) (Robinson, 1999a, Loeuille et al., 2015a). More studies are clearly necessary, but the position of *Allocephalus* as sister group of Lychnophorinae calls to reevaluate the morphological interpretation of the "capitulum" of *Albertinia* and indicates further directions to study the evolution in syncephaly in Lychnophorinae.

The clade grouping *Chronopappus*, *Heterocoma*, *Maschalostachys*, and *Paralychnophora* was also recovered by Loeuille et al. (2015b) but only in one analysis (Bayesian analysis without morphological data). However, it appeared as the sister group of the *Prestelia* Alliance clade (*E. crotonoides* + *Hololepis*) in that study, instead of sister to the derived Lychnophorinae genera, as seen in the present analysis. Similarly to previous phylogenies (Loeuille et al., 2015a, Loeuille et al., 2015b), *Minasia*, *Lychnophorella*, *Piptolepis*, *Lychnophora,* and *Eremanthus* are grouped in a large clade, but its internal relationships vary between the analyses.

Our work did not sample Piptocarphinae, a mainly South American subtribe that includes more than 50 species. Loeuille's work (2015a) shows that the subtribe has affinities with Vernoniinae, Lepidaploinae, and Elephantopinae, although without resolution, indicating this might be a crucial group to help resolving the relationships in the South American clade. Also, as shown by Keeley et al. (2007), the relationships in the African clade are complex, especially close to the transition to South America and should be further investigated with additional sampling, which might help to solve the position of *Stokesia* in relation to the Old and New World clades.

CONCLUSIONS

The Hyb-Seq method used to obtain sequence data for phylogenetic reconstruction proved useful and powerful, allowing us to recover well-resolved and supported relationships in Vernonieae. We consistently recovered the same overall topology regardless of dataset and analysis method, even with incongruence among gene and species trees, with most of the effect of reducing the dataset being the overall decline in statistical support in the tree. Also, we demonstrated the non-monophyly of several subtribes, indicating that further phylogenetic and taxonomic work should be conducted, and that the circumscription of tribe Moquinieae and genus *Distephanus* should be probably reevaluated in relation to their affinity with Vernonieae. The presence of more than 50 monotypic genera in Vernonieae (Keeley and Robinson, 2009) complicates phylogenetic studies, making the sampling process very challenging and possibly indicating an evolutionary history of multiple speciation and extinction events. On the other hand, more complete sampling in future studies may reveal strongly supported clades that could eventually allow a reduction of the number of monotypic genera recognized in the tribe. While the recently developed Hyb-Seq method proved to be reliable, further investigation into Vernonieae phylogeny should focus in improving sampling, especially in lineages that are isolated or morphologically anomalous.

DATA AVAILABILITY STATEMENT

Raw data in the fastq format are deposited at the NCBI Sequence Read Archive, under BioProjects PRJNA540287 and PRJNA546287.

AUTHOR CONTRIBUTIONS

CS designed the study, collected samples, conducted lab work, conducted data analyses, and wrote the manuscript. BL helped designed the study, participated in field collections, provided samples, helped write the discussion section, and reviewed and commented on the manuscript. VF provided samples and reviewed and commented on the manuscript. JM provided samples, collaborated on lab work, data analyses, and reviewed and commented on the manuscript. JP helped design the study and commented and reviewed the manuscript, besides being the thesis advisor for CS's doctoral dissertation.

FUNDING

This work was funded by the Fundação de Amparo a Pesquisa do Estado da São Paulo doctoral scholarships 2013/18189-2 and 2016/12446-1 and by the National Science Foundation Division of Environmental Biology, grant DEB-1745197.

ACKNOWLEDGMENTS

The authors thank ICMBio and IEF-MG for collection permits; the Department of Biosciences and the Center for Biodiversity at the University of Memphis and the Laboratório de Sistemática Vegetal do Departamento de Botânica do IB-USP for material support for the conduct of the laboratorial work, as well as all researchers that provided samples and that were involved in field work. CS thanks Erika Moore for helping design **Figure 7**, Gustavo Heiden and Caetano Oliveira for helping financing the sequencing needed for this work, and Ramhari Thapa, Darrell Brandon, Adam Ramsey, and Mike Ballou for valuable discussion and insights during the development of the paper. This paper is part of CS's doctoral dissertation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01224/full#supplementary-material

SUPPLEMENTARY TABLE 1 | Voucher list of taxa sampled in the study, containing collection data and first date of publication of each sequence.

SUPPLEMENTARY TABLE 2 | General information about the matrices used for phylogenetic analyses, number and identity of loci recovered and molecular evolution models used in gene tree assembly.

REFERENCES

Bankevich, A., Nurk, S., Antipov, D., Gurevich, 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 (5), 455–477. doi: 10.1089/cmb.2012.0021

Barker, M. S., Kane, N. C., Matvienko, M., Kozik, A., Michelmore, R. W., Knapp, S. J., et al. (2008). Multiple paleopolyploidizations during the evolution of the Asteraceae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* 25 (11), 2445–2455. doi: 10.1093/molbev/msn187

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 (7), 1203–1211. doi: 10.3732/ajb.1600113

Blake, S. F. (1945). *Dipterocypsela*, a new genus of Vernonieae from Colombia. *J. Wash. Acad. Sci.* 35 (2), 36–38.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic, A flexible trimmer for Illumina Sequence Data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4, e1660. doi: 10.7717/peerj.1660

Bremer, K. (1994). *Asteraceae: Cladistics & Classification*. Portland: Timber Press.

Bringel, J. B. A., Jr., Nakajima, J. N., and Robinson, H. (2011). *Allocephalus gamolepis*, a new genus and species of Dipterocypselinae (Vernonieae, Asteraceae) from Central Brazil. *Syst. Bot.* 36 (3), 785–788. doi: 10.1600/036364411X583736

Brown, J. W., Walker, J. F., and Smith, S. A. (2017). Phyx: phylogenetic tools for unix. *Bioinformatics* 33 (12), 1886–1888. doi: 10.1093/bioinformatics/btx063

Candolle, A. P. (1836). "Vernoniaceae," in *Prodromus Systematis Naturalis Regni Vegetabilis*, vol. v. 5 . Ed. A. P. de Candolle (Paris: Masson, Paris: Treutel and Würtz), 9–103.

Cassini, H. (1819). Suite du sixième mémoire sur la famille des Synanthérées, contenant les caractères des tribus. *J. Phys. Chim. Hist. Nat. Arts* 88, 189–204.

Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., et al. (2012). Targeted enrichment strategies for next-generation plant biology. *Am. J. Bot.* 99 (2), 291–311. doi: 10.3732/ajb.1100356

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2, more models, new heuristics and parallel computing. *Nat. Methods* 9 (8), 772. doi: 10.1038/nmeth.2109

Dillon, M. O., Funk, V. A., Robinson, H., and Chan, R. (2009). "Liabeae," in *Systematics, Evolution, and Biogeography of Compositae*. Eds. V. A. Funk, A. Susanna, T. F. Stussey, and R. J. Bayer (Vienna, Austria: International Association for Plant Taxonomy (IAPT)), 417–437.

Donoghue, M. J., Doyle, J. A., Gauthier, J., Kluge, A. G., and Rowe, T. (1989). The importance of fossils in phylogeny reconstruction. *Annu. Rev. Ecol. Evol. Syst.* 20, 431–460. doi: 10.1146/annurev.es.20.110189.002243

Edwards, S. V., Xi, Z., Jnake, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., et al. (2016). Implementing and testing the multispecies coalescent module: a valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* 94, 447–462. doi: 10.1016/j.ympev.2015.10.027

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61 (5), 717–726. doi: 10.1093/sysbio/sys004

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32, 786–788. doi: 10.1093/bioinformatics/btv646

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 (8), 1500039. doi: 10.3732/apps.1500039

Funk, V. A., Anderberg, A. A., Baldwin, B. G., Bayer, R. J., Bonifacino, J. M., Breitwieser, I., et al. (2009). "Compositae metatrees: The next generation," in *Systematics, Evolution, and Biogeography of Compositae*. Eds. V. A. Funk, A. Susanna, T. F. Stussey, and R. J. Bayer (Vienna, Austria: International Association for Plant Taxonomy (IAPT)), 747–777.

Funk, V. A., and Chan, R. (2009). "Introduction to Cichorioideae," in *Systematics, Evolution, and Biogeography of Compositae*. Eds. V. A. Funk, A. Susanna, T. F. Stussey, and R. J. Bayer (Vienna, Austria: International Association for Plant Taxonomy (IAPT)), 335–342.

Gamerro, J. C. (1990). Identidad de *Pseudostifftia* con *Moquinia* (Compositae) y consideraciones sobre la ubicación tribal del taxon. *Darwiniana* 30, 123–136.

Grover, C. E., Salmon, A., and Wendel, J. F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis. *Am. J. Bot.* 99 (2), 312–319. doi: 10.3732/ajb.1100323

Guindon, S., and Gascuel, O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Syst. Biol.* 52, 696–704. doi: 10.1080/10635150390235520

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

Huang, C. H., Zhang, C., Liu, M., Gao, T., Qi, J., and Ma, H. (2016). Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Mol. Biol. Evol.* 33 (11), 2820–2835. doi: 10.1093/molbev/msw157

Huang, H., and Lacey Knowles, L. (2016). Unforeseen consequences of excluding missing data from next-generation sequences, simulation studies of RAD sequences. *Syst. Biol.* 65 (3), 357–365. doi: 10.1093/sysbio/syu046

Jansen, R. K., and Palmer, J. D. (1988). Phylogenetic implications of chloroplast DNA restriction site variation in the Mutisieae (Asteraceae). *Am. J Bot.* 75, 751–764. doi: 10.1002/j.1537-2197.1988.tb13496.x

Jeffrey, C. (1988). The Vernonieae of east tropical Africa. Notes on the Compositae: V. *Kew Bull.* 43, 195–277. doi: 10.2307/4113734

Jones, S. B. (1979). Synopsis and pollen morphology of *Vernonia* (Compositae, Vernonieae) in the New World. *Rhodora* 8, 425–447.

Jones, S. B. (1981). Synoptic classification and pollen morphology of *Vernonia* (Compositae: Vernonieae) in the Old World. *Rhodora* 83, 59–75.

Johnson, M. (2017). https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts.

Johnson, M., 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, 695–699. doi: 10.1093/sysbio/syy086

Karis, P. O., Funk, V. A., McKenzie, R. J., Barker, N. P., and Chan, R. (2009). "Arctotideae," in *Systematics, Evolution, and Biogeography of Compositae*. Eds. V. A. Funk, A. Susanna, T. F. Stussey, and R. J. Bayer (Vienna, Austria: International Association for Plant Taxonomy (IAPT)), 385–410.

Keeley, S. C., and Turner, B. L. (1990). A preliminary cladistic analysis of the genus *Vernonia* (Vernonieae: Asteraceae). *Plant Syst. Evol.* 4, 45–66. doi: 10.1007/978-3-7091-6928-5_3

Keeley, S. C., Forsman, Z. H., and Chan, R. (2007). A phylogeny of the "evil tribe" (Vernonieae, Compositae) reveals Old/New World long distance dispersal: support from separate and combined congruent datasets (trnL-F, ndhF, ITS). *Mol. Phylogenet. Evol.* 44, 89–103. doi: 10.1016/j.ympev.2006.12.024

Keeley, S. C., and Robinson, H. (2009). "Vernonieae," in *Systematics, Evolution, and Biogeography of Compositae*. Eds. V. A. Funk, A. Susanna, T. F. Stussey, and R. J. Bayer (Vienna, Austria: International Association for Plant Taxonomy (IAPT)), 439–469.

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 (6), 1695–1701. doi: 10.1093/molbev/mss020

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., and Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol. Biol.* 14 (1), 82. doi: 10.1186/1471-2148-14-82

Liu, L., Xi, Z., Wu, S., Davis, C. C., and Edwards, S. V. (2015). Estimating phylogenetic trees from genome-scale data. *Ann. N. Y. Acad. Sci.* 1360, 36–53. doi: 10.1111/nyas.12747

Loeuille, B., Keeley, S. C., and Pirani, J. R. (2015a). Systematics and evolution of syncephaly in American Vernonieae (Asteraceae) with emphasis on the Brazilian subtribe Lychnophorinae. *Syst. Bot.* 40 (1), 286–298. doi: 10.1600/036364415X686576

Loeuille, B., Semir, J., Lohmann, L. G., and Pirani, J. R. (2015b). A phylogenetic analysis of Lychnophorinae (Asteraceae, Vernonieae) based on molecular and morphological data. *Syst. Bot.* 40 (1), 299–315. doi: 10.1600/036364415X686585

Loeuille, B., Semir, J., and Pirani, J. R. (2019). A synopsis of Lychnophorinae (Asteraceae: Vernonieae). *Phytotaxa* 398, 1–139. doi: 10.11646/phytotaxa.398.1.1

Mai, U., and Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* 19 (S5), 272. doi: 10.1186/s12864-018-4620-2

Mandel, J. R., Dikow, R. B., Funk, V. A., Masalia, R. R., Evan Staton, S., 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 (2), 1300085. doi: 10.3732/apps.1300085

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 (4), 405–410. doi: 10.1111/jse.12265

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.* 116 (28), 14083–14088. doi: 10.1073/pnas.1903871116

Mitchell, N., Lewis, P. O., Lemmon, E. M., Lemmon, A. R., and Holsinger, K. E. (2017). Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. *Am. J. Bot.* 104, 102–115. doi: 10.3732/ajb.1600227

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

Panero, J. L., and Funk, V. A. (2008). The value of sampling anomalous taxa in phylogenetic studies: major clades of the Asteraceae revealed. *Mol. Phylogenet. Evol.* 47, 757–782. doi: 10.1016/j.ympev.2008.02.011

Robinson, H. (1992). Notes on the Lychnophorinae from Minas Gerais, Brazil, a synopsis of *Lychnophoriopsis* Schultz-Bip., and the new genera *Anteremanthus* and *Minasia* (Vernonieae, Asteraceae). *Proc. Biol. Soc. Wash.* 105, 640–652.

Robinson, H. (1994). Notes on the tribes Eremothamneae, Gundelieae and Moquinieae, with comparisons of their pollen. *Taxon* 43, 33–44. doi: 10.2307/1223458

Robinson, H. (1999a). Generic and subtribal classification of American Vernonieae. *Smithsonian Contr. Bot.* 89, 1–116. doi: 10.5479/si.0081024X.89

Robinson, H. (1999b). Revisions in paleotropical Vernonieae (Asteraceae). *Proc. Biol. Soc. Wash.* 112, 220–247.

Roch, S., Nute, M., and Warnow, T. (2019). Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. *Syst. Biol.* 68, 281–297. doi: 10.1093/sysbio/syy061

Schultz-Bipontinus, C. H. (1861). Cassiniaceae uniflorae, oder Verzeichniss der Cassiniaceen mit 1-blüthigen Köpfchen. *Jahresbericht der Pollichia* 18/19, 157–190.

Schultz-Bipontinus, C. H. (1863). *Lychnophora* Martius! und einige benachbarte Gattungen. *Jahresbericht der Pollichia* 20/21, 321–439.

Semple, J. C., and Watanabe, K. (2009). "A review of chromosome numbers in Asteraceae with hypotheses on chromosomal base number evolution," in *Systematics, Evolution, and Biogeography of Compositae*. Eds. V. A. Funk, A. Susanna, T. F. Stussey, and R. J. Bayer (Vienna, Austria: International Association for Plant Taxonomy (IAPT)), 61–72.

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

Stamatakis, A. (2006). RAxML-VI-HPC, maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Stamatakis, A. (2014). RAxML Version 8, A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* 30 (9), 1312–1313. doi: 10.1093/bioinformatics/btu033

Swofford, D. L. (2003). PAUP*: phylogenetic analysis using parsimony, version 4.0 b10.

de Tournefort, J. P. (1700). *Institutiones Rei Herbariae* Vol. 3 vols. Paris: Typographia Regia. doi: 10.5962/bhl.title.713

Vaillant, S. (1719-1723). Établissement de nouveaux caractères de trois familles ou classes de plantes à fleurs composées; sçavoir, des Cynarocéphales, des Corymbifères, et des Cichoracées. *Mémoires de l'Académie Royale des Sciences*.

Wen, J., Egan, A. N., Dikow, R. B., and Zimmer, E. A. (2015). "Utility of transcriptome sequencing for phylogenetic inference and character evolution," in *Next Generation Sequencing in Plant Systematics*. Eds. E. Hörandl and M. S. Apperlhans (Vienna, Austria: International Association for Plant Taxonomy (IAPT)).

Wiens, J. J. (2003). Missing data, incomplete taxa, and phylogenetic accuracy. *Syst. Biol.* 52 (4), 528–538. doi: 10.1080/10635150390218330

Wiens, J. J., and Morrill, M. C. (2011). Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst. Biol.* 60 (5), 719–731. doi: 10.1093/sysbio/syr025

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 (6), 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: 11 November 2019
doi: 10.3389/fpls.2019.01416

Maximize Resolution or Minimize Error? Using Genotyping-By-Sequencing to Investigate the Recent Diversification of *Helianthemum* (Cistaceae)

Sara Martín-Hernanz[1], Abelardo Aparicio[1], Mario Fernández-Mazuecos[2], Encarnación Rubio[1], J. Alfredo Reyes-Betancort[3], Arnoldo Santos-Guerra[3], María Olangua-Corral[4] and Rafael G. Albaladejo[1]*

[1] Departamento de Biología Vegetal y Ecología, Universidad de Sevilla, Sevilla, Spain, [2] Departamento de Biodiversidad y Conservación, Real Jardín Botánico (RJB-CSIC), Madrid, Spain, [3] Jardín de Aclimatación de la Orotava, Instituto Canario de Investigaciones Agrarias (ICIA), Santa Cruz de Tenerife, Spain, [4] Departamento de Biología Reproductiva y Micro-morfología, Jardín Botánico Canario 'Viera y Clavijo'—Unidad Asociada CSIC (Cabildo de Gran Canaria), Las Palmas de Gran Canaria, Spain

OPEN ACCESS

Edited by:
Juan Viruel, Royal Botanic Gardens, Kew, United Kingdom

Reviewed by:
Carolina Granados Mendoza, National Autonomous University of Mexico, Mexico
Julissa Roncal, Memorial University of Newfoundland, Canada
Natascha D. Wagner, University of Göttingen, Germany

***Correspondence:**
Sara Martín-Hernanz
sara.martin.hernanz@gmail.com

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

Received: *16 July 2019*
Accepted: *11 October 2019*
Published: *11 November 2019*

Citation:
Martín-Hernanz S, Aparicio A, Fernández-Mazuecos M, Rubio E, Reyes-Betancort JA, Santos-Guerra A, Olangua-Corral M and Albaladejo RG (2019) Maximize Resolution or Minimize Error? Using Genotyping-By-Sequencing to Investigate the Recent Diversification of Helianthemum (Cistaceae). Front. Plant Sci. 10:1416. doi: 10.3389/fpls.2019.01416

A robust phylogenetic framework, in terms of extensive geographical and taxonomic sampling, well-resolved species relationships and high certainty of tree topologies and branch length estimations, is critical in the study of macroevolutionary patterns. Whereas Sanger sequencing-based methods usually recover insufficient phylogenetic signal, especially in recently diversified lineages, reduced-representation sequencing methods tend to provide well-supported phylogenetic relationships, but usually entail remarkable bioinformatic challenges due to the inherent trade-off between the number of SNPs and the magnitude of associated error rates. The genus *Helianthemum* (Cistaceae) is a species-rich and taxonomically complex Palearctic group of plants that diversified mainly since the Upper Miocene. It is a challenging case study since previous attempts using Sanger sequencing were unable to resolve the intrageneric phylogenetic relationships. Aiming to obtain a robust phylogenetic reconstruction based on genotyping-by-sequencing (GBS), we established a rigorous methodological workflow in which we i) explored how variable settings during dataset assembly have an impact on error rates and on the degree of resolution under concatenation and coalescent approaches, ii) assessed the effect of two extreme parameter configurations (minimizing error rates vs. maximizing phylogenetic resolution) on tree topology and branch lengths, and iii) evaluated the effects of these two configurations on estimates of divergence times and diversification rates. Our analyses produced highly supported topologically congruent phylogenetic trees for both configurations. However, minimizing error rates did produce more reliable branch lengths, critically affecting the accuracy of downstream analyses (i.e. divergence times and diversification rates). In addition to recommending a revision of intrageneric systematics, our results enabled us to identify three highly diversified lineages in *Helianthemum* in contrasting geographical areas and ecological conditions, which started radiating in the Upper Miocene.

Keywords: branch length, diversification, evolutionary radiation, genotyping-by-sequencing, *Helianthemum*, phylogenetic resolution, phylogenomics

INTRODUCTION

The establishment of a robust phylogenetic framework is the initial step for the study of macroevolutionary patterns of specific lineages and requires extensive geographical and taxonomic representativeness, strong statistical support for species relationships and accurate estimates of tree topology and branch lengths. Usually, these goals cannot be achieved in phylogenetic analyses of recently diversified lineages when Sanger sequencing approaches are used. Such techniques typically rely on a small set of relatively slowly evolving loci, which frequently provide insufficient synapomorphies for resolving species relationships. Furthermore, with a small number of loci it is difficult to deal with inconsistencies related to incomplete lineage sorting (ILS; DeFilippis and Moore, 2000; Whitfield and Kjer, 2008) and inter-specific gene flow (Shaw, 2002). As a result, poor resolution and low statistical support are often obtained (DeFilippis and Moore, 2000).

Alternatively, reduced-representation sequencing methods such as restriction-site associated DNA sequencing (RADseq; Miller et al., 2007; Baird et al., 2008; Rowe et al., 2011) and genotyping-by-sequencing (GBS; Elshire et al., 2011) have been shown to be highly efficient in phylogenetic reconstructions of recently diversified lineages given that they allow for the discovery of thousands of genetic markers in non-model species (e.g. Nadeau et al., 2013; Wagner et al., 2013; Fernández-Mazuecos et al., 2018). However, these methods based on Next-Generation Sequencing (NGS) present notable methodological challenges that include i) the high DNA quality generally required (Andrews et al., 2016), ii) the complexity of the assembly and bioinformatic processing (Shafer et al., 2017), iii) the constraints and assumptions of the two approaches currently used in phylogenomics (i.e. concatenation and coalescent approaches; Meiklejohn et al., 2016), iv) the limits of available computing power (Glor, 2010), and v) the biological limitations on data collection (i.e. allele dropout because of mutations at restriction sites; Andrews et al., 2016; **Table S1**).

The assembly and bioinformatic processing of data derived from reduced-representation sequencing methods require many steps and decisions to convert data into a format ready for analysis, which can entail a trade-off between the numbers of loci and SNPs (single-nucleotide polymorphisms) recovered and the magnitude of associated error rates, especially when studying recently diversified lineages (Mastretta-Yanes et al., 2015; Anderson et al., 2017; Lee et al., 2018). Non-optimized values of key assembly parameters such as the clustering threshold, minimum sample coverage and minimum taxon coverage may lead to errors in genotyping and large amounts of missing data (Mastretta-Yanes et al., 2015; Anderson et al., 2017; see **Table S1**), which, in turn, may have an unpredictable impact on phylogenetic inferences in terms of degree of resolution, topology, and branch length estimation (Lemmon et al., 2009; Roure et al., 2013; Mastretta-Yanes et al., 2015; Darriba et al., 2016; Anderson et al., 2017). Furthermore, concatenation and coalescent approaches, frequently used in phylogenomics, are also prone to a number of sources of error that need to be taken into account when reduced-representation sequencing data are used. The concatenation approach, in which all gene alignments are concatenated into a single matrix assuming that all trees share the same history (e.g. Nadeau et al., 2013; Wagner et al., 2013; Cruaud et al., 2014), has been shown to be robust for phylogenetic inference from reduced-representation sequencing data by certain simulations (Rivers et al., 2016). However, other studies indicate that the resulting trees can be misleading in terms of species relationships and tree support (e.g. strong bootstrap support for incorrect relationships) (Kubatko and Degnan, 2007; McVay and Carstens, 2013; **Table S1**) and that this approach is unable to address the problem of ILS (Kubatko and Degnan, 2007). Conversely, the coalescent approach is capable of dealing with ILS and can also be used for constructing species trees in large-scale phylogenomic studies. Within this approach, there are several families of methods, including "summary methods," in which all genes are analysed separately and the resulting gene tree topologies are subsequently or simultaneously used to construct a species tree based on coalescent theory (Liu and Yu, 2011); and "site-based methods," which do not try to estimate gene trees but estimate the species tree directly from the observed site pattern frequencies using properties of the multispecies coalescent model (Chifman and Kubatko, 2014; Vachaspati and Warnow, 2018). Nonetheless, summary methods are sensitive to errors in gene tree estimation (Dupuis et al., 2017) due to insufficient variable sites per locus, and both families of methods may be computationally intensive (reviewed by Liu et al., 2015; Solís-Lemus and Ané, 2016). In general, the limits of available computing power have led researchers to focus on estimating phylogenies of small clades when using reduced-representation sequencing methods (e.g. Jones et al., 2013, Nadeau et al., 2013, Anderson et al., 2017). Taxon-rich clades have been addressed less frequently, even though sampling more taxa affords a wider comparative framework needed for downstream analyses of evolutionary patterns (e.g. divergence time estimates, diversification rate calculations; Hughes et al., 2015; Eaton et al., 2017).

Despite being a challenging case from both systematic and evolutionary standpoints, the genus *Helianthemum* Mill. (Cistaceae) is suitable for testing the trade-off between phylogenetic information and error rates under the two described phylogenomic approaches. *Helianthemum* is by far the largest genus in the Cistaceae, constituting a monophyletic, complex and species-rich Palearctic plant clade with *c.* 140 taxa (104 species and 36 subspecies). Its diversification has probably been driven by the major palaeoclimatic events that have affected the Mediterranean Basin since the Upper Miocene (i.e. the Messinian salinity crisis, the infilling of the Mediterranean Basin and the climatic cycles during the Pleistocene; Aparicio et al., 2017). Despite high geographical and taxonomical representativeness, a previous attempt to infer phylogenetic relationships in *Helianthemum* based on Sanger sequencing of combined ITS and cpDNA sequences (Aparicio et al., 2017) resulted in very low resolution and low statistical support for shallow nodes. However, support was recovered for three main clades with intriguing systematic and evolutionary patterns. In particular, the internal topologies of these three clades were similar, each including a species-rich subclade (corresponding with the three largest taxonomical sects. *Eriocarpum*, *Pseudocistus,* and *Helianthemum*) sister to poorly

diversified subclades, an asymmetry that can be an indicator of recent and rapid radiations (Nee et al., 1996; Sanderson and Donoghue, 1996; Pybus and Harvey, 2000).

The main aim of this study was to generate a robust species and subspecies-level phylogenetic reconstruction of the genus *Helianthemum* based on the analysis of paired-end GBS data. For this purpose, we conducted an extensive geographical and taxonomic sampling, including over 70% of the species and subspecies of *Helianthemum*, and representing all the supraspecific taxa (2 subgenera, 10 sections). Thus, our study provides the most comprehensive phylogenetic hypothesis for the genus *Helianthemum* and one of the largest trees reconstructed to date based on reduced-representation sequencing (e.g. Wagner et al., 2013; Ebel et al., 2015). This phylogeny was generated by following a rigorous methodological workflow (see **Figure 1**) in which we aimed to i) explore how bioinformatic decisions affect error rates (locus, allele and SNP error) and degree of resolution in phylogenetic inferences using concatenation and coalescent approaches; ii) assess the effects of two extreme configurations of assembly parameters (minimizing error rates vs. maximizing phylogenetic resolution) on tree topology and branch length estimation; and iii) evaluate the effects of these configurations on estimates of divergence times and diversification rates.

The robust phylogenetic framework here established provides, for the first time, the opportunity to address questions about the macroevolutionary patterns of the genus *Helianthemum*. Specifically, we tested if the large number of species and subspecies in the genus is the result of low extinction rates or, conversely, of recent and rapid independent radiations corresponding with the three largest sections. With the powerful insights provided by the molecular phylogenies comes the possibility of detecting rapid and recent radiations in particular groups based on three operational criteria: i) a recent common ancestor, ii) species-poor sister lineages, and iii) significant bursts of diversification (Nee et al., 1996; Sanderson and Donoghue, 1996; Pybus and Harvey, 2000; Schluter, 2000; Glor, 2010; Bouchenak-Khelladi et al., 2015). Since the recent common ancestry of each of the three largest sections of *Helianthemum*, as well as diversity asymmetries with their sister clades have already been suggested (Aparicio et al., 2017), here we aim to explore if significant bursts of diversification are detectable during the evolutionary history of the genus. In this regard, we asked: i) How high is the diversification rate in *Helianthemum* and in the three largest sections compared to other recently diversified Mediterranean lineages? ii) Is there any detectable acceleration of diversification rates in the course of *Helianthemum* evolution? If so, iii) do these accelerations correspond with the origin of the three largest sections and thus provide additional evidence of recent and rapid radiations? And iv) are these alleged independent radiations characterised by contrasting diversification patterns?

MATERIALS AND METHODS

Taxon Sampling

One hundred and twenty-eight samples were used in this study (**Table S2**). The ingroup consisted of 98 taxa (73 species, 25 subspecies; 124 accessions; **Tables S2 and S3**) from the whole distribution range of the genus *Helianthemum*, including all supraspecific taxonomic ranks (2 subgenera, 10 sections). Given the large geographical and taxonomic scope, all species and subspecies were represented by a single sample each, except those belonging to monospecific or species-poor sections and those not included in the previous phylogenetic reconstruction of the genus (Aparicio et al., 2017), for which two samples were included. Replicates from three individual samples representing the three main lineages of *Helianthemum* (Aparicio et al., 2017; **Table S2**) were also included to optimize bioinformatic processing (see *Materials and Methods*, *Bioinformatics Workflow*). The outgroup consisted of four species belonging to other genera of Cistaceae, one representing an early-diverging lineage within the family (*Fumana*) and the other three (*Cistus*, *Halimium* and *Tuberaria*) representing the well-supported sister clade to *Helianthemum* (Aparicio et al., 2017). The inclusion of this outgroup enabled the implementation of two of the three fossil calibration points in the dating analysis (see *Materials and Methods*, *Downstream Analyses*). Except for four samples obtained from herbarium collections, all the plant material used in this study was freshly collected in the field from natural populations and stored in silica gel until DNA extraction (**Table S2**).

DNA Extraction, Library Preparation and NGS

DNA was extracted from the silica-dried leaf material using the Bioline Isolate II Plant DNA Kit (Bioline, London, UK) following the manufacturer's protocol. The concentration and quality of each sample were assessed using a Qubit dsDNA BR Assay kit (Thermo Fisher Scientific), and 260/280 and 260/230 absorbance ratios were measured on a NanoDrop spectrophotometer (Thermo Fisher Scientific). Paired-end genotyping-by-sequencing (PE GBS) multiplexed libraries were constructed and sequenced by CNAG (Centro de Análisis Genómicos, Barcelona, Spain) following the protocol used by Elshire et al. (2011) with improvements from Poland et al. (2012) and Sonah et al. (2013). The restriction enzyme *Ape*K1 was chosen for digestion of genomic DNA based on a small-scale experiment. Two lanes of Illumina HiSeq 2000, with a read length of 2x125bp, were used to increase sequencing coverage. Image analysis, base calling and quality scoring of the run were conducted using the manufacturer's software Real Time Analysis (RTA 1.18.66.3), followed by generation of FASTQ sequence files by CASSAVA (see **Methods S1** for details).

Bioinformatics Workflow

Due to the complexity of the proposed methodology, which contains three main steps (exploratory PyRAD assembly, final PyRAD assembly and downstream analyses) and several analyses within each one (error rate calculations, concatenated and coalescent phylogenetic analyses, branch length estimation, divergence time estimation and diversification rate analyses), the bioinformatics and analytical workflow followed in this study is summarized in **Figure 1**, based on Anderson et al. (2017).

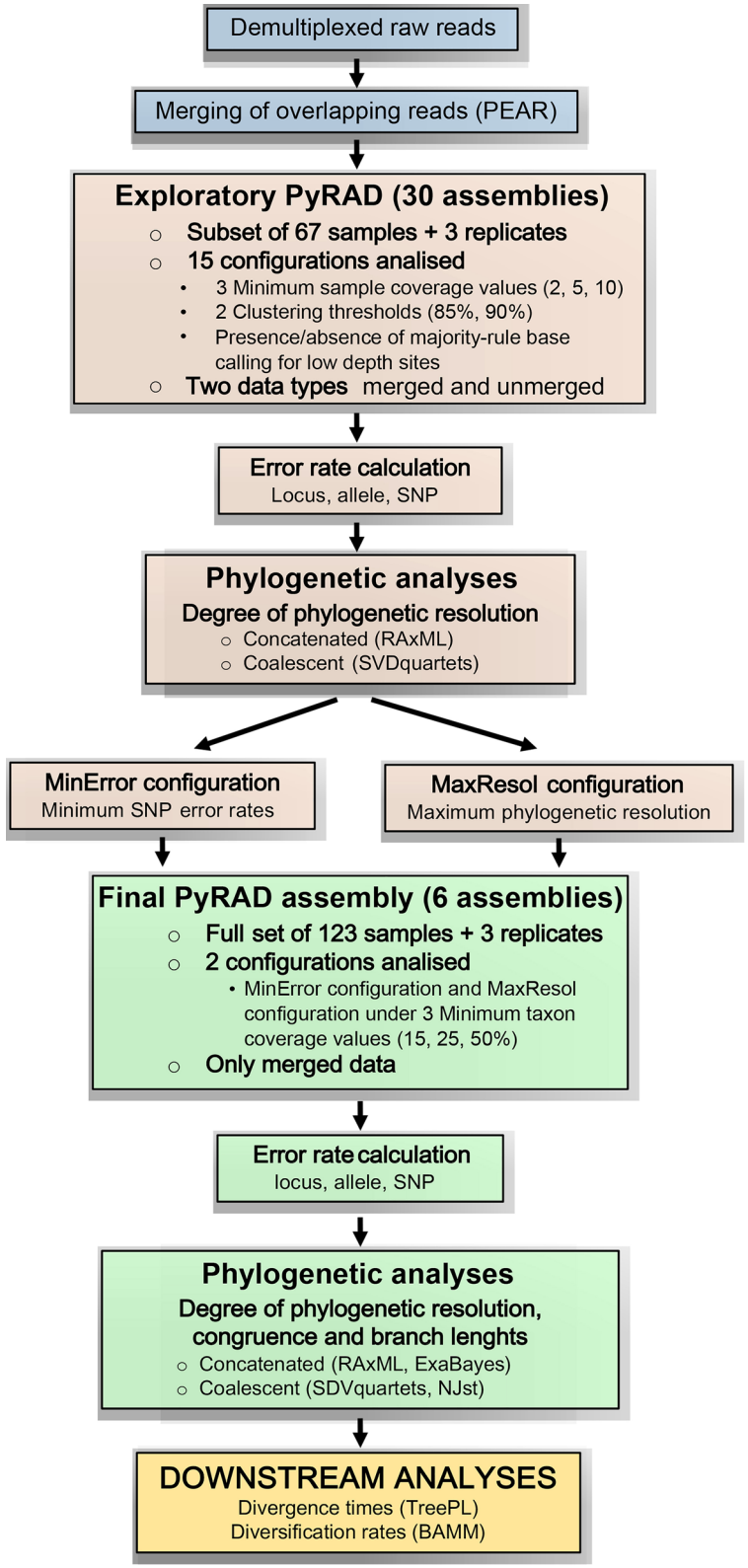

FIGURE 1 | Bioinformatic and analytical workflow used to process genotyping-by-sequencing (GBS) data for the genus *Helianthemum* (modified from Anderson et al., 2017). Blue rectangles represent the pre-processing of raw reads applied to all studied samples; brown rectangles represent the exploratory analyses applied to a subset of the studied samples; green rectangles represent the final analyses applied to the full set of studied samples; and the yellow rectangle represents the downstream analyses also applied to the full set of studied samples.

Demultiplexing and Merging of Overlapping Reads

Demultiplexing was carried out using a custom script developed by CNAG in which GBS and Illumina barcodes as well as reads shorter than 25 bases were removed. The demultiplexed Illumina FASTQ reads were run on PEAR v. 0.9.8 (Zhang et al., 2014) to check for and merge overlapping reads using default settings except 33 bp as the minimum possible length of the assembled sequences (-n option) and 33 bp as the minimum length of reads after trimming the low quality part (-t option). Merging the reads is advisable to reduce duplication in the dataset and increase the reliability of each nucleotide position, especially at the ends of the reads which tend to have higher error rates (Eaton, 2014; Andrews et al., 2016; Anderson et al., 2017).

Exploratory PyRAD Assembly

Reads were assembled *de novo* using the PyRAD pipeline v. 3.0.6 (Eaton, 2014) since no reference genome was available for the family Cistaceae. Before the assembly, a quality filtering step was run in which bases with a FASTQ quality score below 20 were replaced with N and sequences having more than 4% of Ns were discarded. Merged and unmerged output files generated by PEAR were assembled and analysed separately by setting the data type to "merged" or "pairend" respectively in the PyRAD parameter file (parameter 11).

To determine the appropriate assembly settings, we followed the approach of Mastretta-Yanes et al. (2015) using replicates to assess the error rates associated with different parameter configurations (three pairs of replicates, six samples in total), as well as the approach used by Anderson et al. (2017) to analyse the impact of different parameter values on the degree of resolution of resulting phylogenetic trees in terms of number of supported nodes (see *Materials and Methods*, *Phylogenetic Analysis*). In particular, Mastretta-Yanes et al.'s approach was built on the idea that individual sample replicates (consisting of two DNA extractions from the same sample that are sequenced, processed and analysed independently), under the expectation of identical genotypes, allow the quantification of genotyping errors as the differences between replicates at the locus, allele, and SNP levels in the absence of a reference genome. Thus, locus error represents the number of loci missing from one replicate but not from the other relative to the total number of loci; allele error is the number of shared loci differing in sequence between the replicates relative to the total number of shared loci; and SNP error is the number of SNPs differing between replicates (hard error when differing in both alleles and heterozygous error when differing in one allele) relative to the total number of shared SNPs. Because replicates derived from the same DNA sample should have the same genotype, one can evaluate which parameter values of the assembly pipeline maximize the number of loci while minimizing differences between replicate pairs (see Appendix S1 from Mastretta-Yanes et al., 2015).

The bioinformatic parameters evaluated were the type of data (merged or unmerged), the clustering threshold, the base calling method (statistical base calling or majority-rule base calling), the minimum sample coverage and the minimum taxon coverage (Eaton, 2014; see **Methods S2** for details). All other parameters were set to default values. To reduce computing time and simultaneously allow a robust evaluation of assembly settings, these exploratory analyses were carried out for a subset of 70 samples representing all suprageneric taxonomic ranks. The subset was run through 15 parameter configurations (30 assemblies in total including merged and unmerged data): a minimum sample coverage of 2, 5, or 10 per individual locus, presence/absence of majority-rule base calling for low depth sites (from a minimum sample coverage below 5 or 10), clustering threshold at 85%, 90%, and a combination of 90% in step 3 (clustering within samples) and 85% in step 6 (clustering among samples). The minimum taxon coverage was kept at 15%. Locus error, allele error, and SNP (hard and heterozygous) error rates were calculated with modified python and R scripts used by Anderson et al. (2017) (scripts 5–7 contained in Supporting Information S3 of that article) and ape v. 3.3. (Paradis et al., 2004) for each of the three replicated samples and then averaged for each configuration.

Final PyRAD Assembly

We selected two extreme parameter configurations to analyse the full set of samples: the first one minimizing allele and SNP error rates (MinError configuration) and the second one maximizing phylogenetic resolution (MaxResol configuration). The latter was defined as the configuration that provided the highest number of supported nodes in phylogenetic analyses (see *Materials and Methods*, *Phylogenetic Analyses*). The resulting MinError configuration had a minimum sample coverage of 10, no majority-rule base calling, a clustering threshold of 90% and was based on merged data. The MaxResol configuration had minimum sample coverage of 10, majority-rule base calling, a clustering threshold of 85% and was based on merged data (**Table 1**). Both configurations were applied to the full set of samples, and outputs were generated at minimum taxon coverage values of 15, 25 and 50% (six assemblies in total; see **Methods S2** for details) to assess the impact of the amount of missing data on the degree of resolution (number of supported nodes), congruence between phylogenetic trees and branch length estimates (see *Materials and Methods*, *Phylogenetic Analyses*).

Phylogenetic Analyses

To analyse the impact of assembly parameters (see *Materials and Methods*, *Bioinformatic workflow*) on phylogenetic resolution, we applied two phylogenetic methods to the subset assemblies resulting from the exploratory PyRAD analyses: a concatenated approach using maximum likelihood (ML) in RAxML 7.2.8 (Stamatakis, 2006) and a coalescent approach using the quartet-based method SVDquartets (Chifman and Kubatko, 2014) implemented in PAUP* 4 (Swofford, 2002). ML analyses were conducted using the GTR+GAMMA nucleotide substitution model. This widely used model was chosen because it usually fits real data better than other simpler alternative models (Sumner et al., 2012). At the same time it is practical for large data sets compared to more complex models (e.g. GMM by Barry and Hartigan, 1987; SBH and RBH models by Jayaswal et al., 2011).

TABLE 1 | Assembly information obtained from the exploratory PyRAD assembly using the subset of 70 samples.

PYRAD PARAMETERS				ASSEMBLY RESULTS										PHYLOGENETIC INFERENCES RESULTS	
Data type	**Majority-rule base calling**	**Minimum sample coverage**	**Clustering threshold**	**Base pairs**	**Number of loci**	**Number of SNPs**	**N° of phylogenetically informative sites**	**% missing data**	**Locus error**	**Allele error**	**SNP error**	**Hard error**	**Het error**	**RaxML resolution**	**SVD quartets resolution**
Merged	No	2	85	2509874	20565	448042	240782	69.10%	**0.0778**	0.1101	0.0053	0.0033	0.002	91.18%	**78.33%**
		2	90	2302252	19212	318202	152422	*70.50%*	0.0737	0.0901	0.0048	0.0024	0.0023	88.24%	70.00%
		2	90_85	2054634	17118	355695	190512	68.90%	0.0803	0.1	0.0044	0.0024	0.002	92.65%	*73.33%*
		5	85	1154519	9793	213781	116633	68.70%	0.1201	0.0852	0.0038	0.0022	0.0016	88.20%	*63.33%*
		5	90	1038637	8982	149295	72971	69.80%	0.1133	0.0657	0.0032	0.0016	0.0015	92.65%	63.33%
		5	90_85	1032374	8859	190199	104427	68.80%	0.1229	0.0792	0.0034	0.0019	0.0016	86.76%	71.67%
		10	85	484189	4210	90658	50295	67.80%	0.2014	0.0609	0.0042	0.003	0.0011	*78.00%*	76.67%
		10	90	*424790*	*3758*	*62176*	53795	68.40%	0.1866	**0.0421**	**0.0025**	**0.0014**	**0.0011**	79.40%	*63.33%*
		10	90_85	461589	4021	86212	*45637*	*67.80%*	*0.2065*	0.054	0.0028	0.001663	0.0011	79.41%	73.77%
	Yes	5	85	2717090	22238	504878	270162	69.10%	0.0808	0.1303	0.0073	0.0047	0.0026	97.06%	73.33%
		5	90	2468923	20496	355049	170002	70.40%	0.0737	0.1086	0.0069	0.0037	0.0032	91.18%	73.33%
		5	90_85	2207222	18313	393910	211306	68.80%	0.0798	0.1204	0.0066	0.0038	0.0028	94.12%	76.67%
		10	85	**3421305**	**28311**	**645213**	**3421305**	**67.40%**	0.0801	*0.1338*	*0.0092*	*0.0064*	0.0028	**100%**	73.33%
		10	90	3169491	26363	462983	230582	69.50%	0.0739	0.115	0.0092	0.0057	*0.0035*	94.12%	71.67%
		10	90_85	2668014	22254	477571	262214	67.50%	0.0803	0.1257	0.0082	0.0054	0.0027	91.18%	71.67%
Unmerged	No	2	85	472248	2135	69171	38676	71.00%	0.1717	0.1359	0.0057	0.0036	0.0021	**80.88%**	**66.67%**
		2	90	227656	1022	21605	11112	70.90%	**0.1494**	0.1186	0.0039	0.0016	0.0023	64.71%	53.33%
		2	90_85	410904	1858	58964	33289	70.70%	0.1713	0.1292	0.0036	0.002	0.0017	73.53%	63.33%
		5	85	59990	271	6613	3642	64.20%	0.1968	0.0714	0.011	0.0073	0.0037	44.10%	25.00%
		5	90	39745	179	2203	1094	59.40%	0.1639	0.0421	0.0027	0.0027	0	39.70%	18.33%
		5	90_85	55234	250	5822	55234	62.00%	0.212	0.0479	0.0038	0.0016	0.0022	47.06%	26.67%
		10	85	22353	101	1002	574	55.40%	0.1617	0.0263	0.0072	0	0.0072	41.50%	10.00%
		10	90	20399	*92*	*566*	*292*	57.10%	0.1558	0.0057	0	0	0	36.80%	*8.33%*
		10	90_85	*20359*	*92*	699	393	**55.10%**	0.1703	**0.0056**	0	0	0	*33.82%*	13.33%
	Yes	5	85	389645	3049	86228	47461	74.30%	0.2823	*0.2729*	0.0351	0.0306	*0.0044*	54.41%	43.33%
		5	90	145568	1120	23778	13160	72.60%	0.2604	0.1386	0.0195	0.0192	0.0003	50.00%	18.33%
		5	90_85	189955	1444	37927	21722	73.10%	0.2595	0.1819	*0.0351*	*0.0346*	0.0005	58.82%	38.24%
		10	85	**573516**	**5186**	**133793**	**573516**	74.30%	0.3011	0.1541	0.0252	0.0251	0.0001	70.58%	46.67%
		10	90	474711	4277	95182	53795	74.30%	*0.3013*	0.1437	0.0182	0.0182	0	63.24%	50.00%
		10	90_85	527721	4766	120850	69793	*74.40%*	0.3	0.1563	0.0221	0.0221	0	66.67%	48.33%

Numbers in italic indicate the worst values and numbers in bold indicate the best ones. The MaxResol configuration corresponds with minimum sample coverage (md) = 10 and majority-rule base calling under this coverage and clustering threshold (ct) = 85 from merged data. The MinError configuration corresponds with md = 10 and ct = 90 from merged data.

We applied a rapid bootstrap with automatic bootstrap stopping criterion and calculation of extended majority-rule consensus tree, followed by search for the best-scoring ML tree. No partition scheme was applied. The quartet-based method SVDquartets was selected given its computational efficiency, which makes it highly suitable for estimation of species trees of large taxon sets. The SVDquartets analysis was run under the multispecies coalescent using the concatenated alignment, evaluating one million quartets. One thousand bootstrap replicates were conducted and results were summarised in a 50% majority-rule consensus tree. After evaluating the degree of resolution provided by merged and unmerged data separately (see details below), we combined both types of data and checked whether this resulted in an improvement in phylogenetic resolution. Since no significant improvement was obtained and given that the error rates were substantially higher for unmerged data (see *Results*), we only analysed merged data for the full set of samples.

We performed the same analyses (RAxML and SVDquartets) for the two selected configurations (MinError and MaxResol) using the full set of samples under three values of the minimum taxon coverage parameter (15%, 25%, and 50%). We implemented an additional concatenated analysis using Bayesian inference (BI) in ExaBayes 1.4.1 (Aberer et al., 2014), as well as a further coalescent-based analysis using the NJst method (Liu and Yu, 2010). BI was implemented with the GTR+GAMMA substitution model and one or two runs (until convergence was reached) with four Metropolis-coupled Monte Carlo Markov Chains (MCMCs) each, and trees sampled every 500 generations for 500 000 generations. Convergence was assessed with Tracer 1.7.1 (Rambaut et al., 2018) using summary statistics calculated from the parameter files. We checked that a minimum value of 200 had been reached for the effective sample sizes (EES) of all parameters. Fifty-percent majority-rule consensus phylograms and posterior probabilities were obtained using the *consense* command with a burn-in fraction of 10%.

Amongst available summary methods accounting for ILS, we selected NJst because it is able to infer the species tree from unrooted gene trees (outgroup samples would be absent from many gene trees in our dataset, impeding the rooting of gene trees) and it can accommodate missing data. To build the species trees under the NJst method, we firstly estimated gene trees using RAxML with the GTR+GAMMA substitution model and 200 bootstrap replicates for all loci showing variability. One hundred multilocus bootstrap replicates (Seo, 2008; Mallo, 2015) were generated, thus resampling nucleotides within loci, as well as loci within the dataset. The NJst method was implemented on the one hundred bootstrapped matrices using the R script NJstM (Mallo, 2016), which relies on the phybase package (Liu and Yu, 2010). A 50% majority-rule consensus tree was then built from the 100 bootstrap replicates in PAUP* 4 (Swofford, 2002).

All phylogenetic analyses and the bioinformatic processing in PyRAD (see *Materials and Methods*, *Bioinformatics Workflow*) were performed using the computer clusters at the Centro Informático Científico de Andalucía (CICA, Seville, Spain) and the Consejo Superior de Investigaciones Científicas (cluster Trueno, CSIC, Madrid, Spain).

We evaluated the degree of resolution in the trees inferred from all parameter configurations (subset and the full set of samples) by calculating the quotient of the number of resolved nodes (bootstrap support BS > 70; posterior probability PP > 0.90; Hillis and Bull, 1993; Salichos et al., 2014), relative to the total number of nodes in the tree. Since traditional branch support metrics (BS, PP) present problems of tractability and interpretation when applied to phylogenomic datasets (Pease et al., 2018), we additionally implemented the recently developed Quartet Sampling (QS) method (Pease et al., 2018) using the MinError and MaxResol Bayesian trees. This method represents a generalized framework to quantify phylogenetic uncertainty (specifically branch support) that distinguishes branches with low information from those with multiple highly supported, but mutually exclusive, phylogenetic histories by calculating three metrics: Quartet Concordance (QC) score, Quartet Differential (QD) score, and Quartet Informativeness (QI) score (Pease et al., 2018). For each analysis, we ran 100 replicates per internal branch. We were most interested in QC, the frequency of quartets sampled that are concordant with the consensus tree.

For the full-set assemblies, we assessed the congruence among trees resulting from the two configurations following two approaches: i) by comparing Bayesian trees from ExaBayes (because of their highest resolution; see *Results*) using the relative Robinson–Foulds (RF) distance (Robinson and Foulds, 1981) and the Kuhner–Felsenstein branch score difference (BSc) (Kuhner and Felsenstein, 1994), calculated with the "RF.dist" and "KF.dist" functions of the R package phangorn v. 2.5.3 (Schliep, 2011); and ii) by visually inspecting incongruent placements of individual samples or whole clades (Pirie, 2015). Finally, we evaluated the potential influence of error rates and proportion of missing data (resulting from the three values of minimum taxon coverage: 15%, 25%, and 50%) on branch length estimates in the RaxML and ExaBayes trees for the full-set assemblies and the two extreme configurations. Thus, for each tree we calculated median values of terminal branch lengths and median values of internal branch lengths divided by the total branch length of the tree (relative branch lengths) using ape v. 3.3 (Paradis et al., 2004). The R package ggplot2 v.3.1.1 (Wickham, 2009) was used to visualize the results.

Downstream Analyses

Divergence Times

Divergence times were estimated using the penalized likelihood (PL) approach implemented in the program TreePL v. 1.0 (Smith and O'Meara, 2012). Penalized likelihood (Sanderson, 2002) uses a tree with branch lengths and age constraints for time calibration without prior parametric distributions. It considers rates to be auto-correlated and further accounts for among-branch rate heterogeneity, using a so-called smoothing parameter (Sanderson, 2002). TreePL is a modified and speed-enhanced version of the program r8s (Sanderson, 2003) using stochastic optimization and hill-climbing gradient-based methods, more suitable for very large data sets. We utilized TreePL because most other approaches for divergence time estimation (e.g. the uncorrelated lognormal relaxed clock approach in BEAST; Drummond et al., 2006;

Drummond and Rambaut, 2007) would not be practical given the large number of taxa and loci analysed here.

We used the phylogenetic trees resulting from ExaBayes as input (except that resulting from the MinError configuration under 50% minimum taxon coverage due to its low resolution). As penalized likelihood does not automatically provide confidence intervals, we conducted the analysis using the majority-rule consensus trees resulting from the Bayesian analyses in ExaBayes (see above) and 900 trees from the Bayesian distribution of the same analyses after a 10% burnin. Trees were pruned to include only one terminal per species. A "priming" analysis was first conducted to optimize the set of parameters. Based on these results, the values of gradient-based, auto-differentiation-based, and auto-differentiation cross-validation-based optimizers were all set to two.

For the implementation of fossil calibration points, PL approaches need either a defined fixed age of a node, or a minimum and/or a maximum age constraint on a node. We applied four minimum and maximum age constraints as calibration points (N1: stem node of genus *Tuberaria*, min = 3.02 Myr, max = 10.53 Myr; N2: stem node of genus *Helianthemum*, min = 7.07 Myr, max = 23.86 Myr; N3: crown node of genus *Helianthemum*, min = 3.56 Myr, max = 14.08 Myr; and N4: stem node of *Helianthemum nummularium* complex, min = 0.32, max = 3.61). The minimum ages used in N1, N2, and N4 are fossil-based age constraints (Naud and Suc, 1975; Menke, 1976; Hrynowiecka and Winter, 2016) while the maximum ages in those calibration points as well as the minimum and maximum ages used in N3 are estimates obtained from a previously-published dated phylogeny of Cistaceae (Aparicio et al., 2017) using BEAST (Drummond et al., 2012).

The analysis was set to be thorough to make sure that it continued to iterate until convergence. We selected a smoothing parameter with values between $1x10^{-199}$ and $1x10^{-9}$ depending on the tree, following the random subsample and replicate cross-validation approach (RSRCV) as implemented in TreePL, in which 235 values from $1x10^{-226}$ to $1x10^{8}$ were tested. RSRCV produces similar results to those using standard cross-validation (i.e. removing one taxon), but is capable of handling trees with thousands of taxa within a reasonable time frame (Smith and O'Meara, 2012). The chronograms resulting from the 900 Bayesian trees were then summarized with TreeAnnotator v1.7.5 (Drummond et al., 2012), and 95% confidence intervals were represented on the chronogram resulting from the majority-rule consensus tree to incorporate topological and branch length uncertainty.

Diversification Rates

First, we estimated absolute net diversification rates for the genus *Helianthemum* and for the three largest sections, and compared them with the most rapid episodes of hyper-diversification reported for other Mediterranean plant lineages (Vargas et al., 2018). We used the standardized whole-clade method of Magallón and Sanderson (2001) implemented in the R package geiger v. 2.0.6.1 (Harmon et al., 2008). Rates were calculated for the mean crown ages obtained from a previously published chronogram (Aparicio et al., 2017) because these ages were estimated using a Bayesian relaxed clock analysis of specific DNA regions obtained by Sanger sequencing, as in most of the other Mediterranean examples used here for comparison.

Secondly, we applied a Bayesian approach implemented in BAMM v. 2.5.0 (Bayesian analysis of macroevolutionary mixtures: Rabosky et al., 2013; Rabosky et al., 2014a; Shi and Rabosky, 2015) to detect significant changes in diversification dynamics (speciation and extinction rates). A significant increase in diversification rate is considered an evidence of the initiation of a radiation (Bouchenak-Khelladi et al., 2015). BAMM uses 'reversible jump' Markov chain Monte Carlo (rjMCMC) to account for rate variation through time and among lineages (Rabosky, 2014). BAMM was applied using both TreePL chronograms and MCMC analyses were run with four chains for $10x10^{6}$ generations, sampling every 5000 generations. To account for the non-random sampling of our data set, we assigned sampling fractions at section level (**Table S3**). The prior distributions on speciation (λ) and extinction (μ) rates were estimated with the R package BAMMTOOLS v. 2.1.0 (Rabosky et al., 2014b) using the 'setBAMMprior' command. Likewise, calculation of ESS for the log-likelihood and the number of shift events, as well as post-run analyses and visualization of results were conducted with BAMMTOOLS. Diversification rate variation among the clades of our *Helianthemum* tree was evaluated with the following approaches: i) mean diversification rates at any point along every branch of the tree were displayed as a phylorate plot, ii) the best overall shift configuration was estimated as the maximum shift credibility (MSC) configuration, which maximizes the marginal probability of rate shifts along individual branches, and iii) speciation rates of the three largest sections were visualized as rate-through-time plots.

RESULTS

Exploratory and Final PyRAD Assemblies

The number of read pairs, the number of merged, unmerged and discarded reads in PEAR and the number of loci recovered in PyRAD for each sample under both parameter configurations are shown in **Table S4**. The total number of loci recovered from the *exploratory PyRAD assembly* using the subset of 70 samples ranged from 3758 to 28311 in merged datasets and from 92 to 5186 in unmerged datasets, demonstrating the dramatic effect of parameter selection on the amount of resulting data (**Table 1**). In particular, the number of SNPs and PIS (phylogenetically informative sites) in the assembly decreased as the minimum sample coverage and clustering threshold increased. The implementation of majority-rule base calling resulted in larger datasets than statistical base calling alone. The recovered error rates based on three replicate samples also varied considerably (**Table 1**). In this case, as minimum sample coverage increased, locus error rates increased and allele and SNP error rates decreased. Furthermore, a similarity threshold of 90% always recovered error rates lower than those obtained under the 85% threshold and under the combination of 90% in step 3, and 85% in step 6. Finally, error rates were always lower in analyses of merged data than in analyses of unmerged data under the same parameter values (**Table 1**).

Regarding the full-set assemblies, the proportion of missing data varied between 33.7%, and 77.1%; fewer missing data were recovered as the minimum taxon coverage increased (**Table 2**). In the same way, the number of SNPs and PIS decreased as the minimum taxon coverage increased, especially from 25% to 50%. Lastly, although locus error increased with increasing minimum taxon coverage, allele and SNP error rates decreased.

Phylogenetic Analyses

Degree of Resolution, Congruence and Branch Length Estimation

Phylogenetic method, data type (merged vs. unmerged), minimum sample coverage and minimum taxon coverage all significantly impacted the degree of resolution of phylogenetic trees (**Tables 1** and **2**). Tree resolution resulting from the concatenated analyses was higher than that obtained from coalescent analyses, especially in sects. *Pseudocistus* and *Helianthemum* (see below), and improved as the amount of data increased. In particular, MaxResol configuration assemblies recovered a higher degree of resolution in most of the analyses than MinError configuration assemblies. In the same way, the minimum taxon coverage parameter had a serious effect on the degree of resolution, particularly for the smallest assembly (MinError configuration, minimum taxon coverage = 50%), in which there was essentially no resolution within the three largest sections of the inferred phylogeny, probably due to a dramatic loss of phylogenetic information (**Table 2**). However, the MinError configuration yielded well-resolved phylogenetic trees under the two concatenation methods when minimum taxon coverage was 15% (RAxML: 79.34%; ExaBayes: 97.52%), which does not differ greatly from the results under the MaxResol configuration (RAxML: 90.00%; ExaBayes: 97.87%) (**Figure S1**). The exceptions were some minor incongruences that were well supported based on BS and PP metrics and mainly involved shallow nodes within sects. *Helianthemum* and *Pseudocisuts* (**Figure 2**). Consistent with these incongruences, the quartet sampling analyses displayed negative QC scores for these conflictive nodes (**Figure 3**). Negative scores imply that one of the discordant topologies is the most commonly resampled quartet. Despite these few topological discordances, QC and QI scores were high for most of the nodes, indicating a generally robust phylogenetic inference in both configurations and a strong topological consensus between them.

Total and mean branch lengths were substantially higher for the MaxResol than for the MinError configuration, and decreased as minimum taxon coverage increased for both configurations (**Table 2**). However, relative internal branch lengths stayed essentially constant across assemblies while relative terminal branch lengths were considerably longer under MaxResol than under MinError (**Figure 4**).

RF distances between assemblies within the MaxResol configuration were lower than within the MinError configuration or between assemblies from different configurations (**Table 3A**). BSc distances, a more appropriate measure in our context (because it takes branch length differences into account), were lower between assemblies within the MaxResol and MinError configurations than between assemblies from different configurations (**Table 3B**).

Overall, tree topology and branch length estimates were more affected by parameter configuration (defined by base calling

TABLE 2 | Characteristics of assembled genotyping-by-sequencing datasets from the *final PyRAD assembly.*

			MaxResol configuration			MinError configuration		
			MinCov15%	**MinCov25%**	**MinCov50%**	**MinCov15%**	**MinCov25%**	**MinCov50%**
Assembly information		Number of bp	3596013	1263524	239766	630754	158884	31706
		Number of loci	30351	10968	2214	5768	1471	295
		Number of SNPs	735769	309885	71477	96241	27130	4191
		Number of PIS	409337	182405	46097	47402	14055	2349
		Number of singleton sites	265805	102808	19809	27865	6954	891
		Percentage of missing data	74.40%	60.30%	34.70%	77.10%	61.10%	33.70%
Error rates		Locus error	0.0718	0.0889	0.1101	0.1450	0.1981	0.1718
		Allele error	0.1274	0.1089	0.0849	0.0408	0.0291	0.0133
		SNP error	0.0086	0.0063	0.0053	0.0022	0.0014	0.0006
		Hard error	0.0062	0.0045	0.0040	0.0012	0.0006	0.0006
		Heterozygous error	0.0024	0.0018	0.0013	0.0011	0.0008	0.0000
Phylogenetic analyses	RAxML	Resolution	95.04%	96.69%	90.08%	79.34%	62.81%	48.76%
		Total branch length	1.8256	1.7033	1.2817	0.7288	0.5218	0.2451
		Mean branch length	0.0073	0.0068	0.0051	0.0029	0.0021	0.0010
	ExaBayes	Resolution	94.21%	100%	98.35%	97.52%	86.78%	52.89%
		Total branch length	1.8197	1.7007	1.2846	0.7311	0.5271	0.2599
		Mean branch length	0.0073	0.0068	0.0051	0.0029	0.0021	0.0010
	SVDquartets	Resolution	77.69%	76.03%	71.70%	58.68%	49.59%	24.79%
	NJst	Resolution	82.65%	79.59%	54.98%	30.93%	26.80%	14.43%

Assembly information obtained from the final PyRAD assemblies using the full set of 126 taxa. Error rates and phylogenetic analysis information were obtained from two extreme parameter configurations (MaxResol, maximizing phylogenetic resolution; and MinError, minimizing error rates) under three minimum taxon coverage percentages (15, 25 and 50%). SNP, single-nucleotide polymorphism. PIS, phylogenetically informative sites.

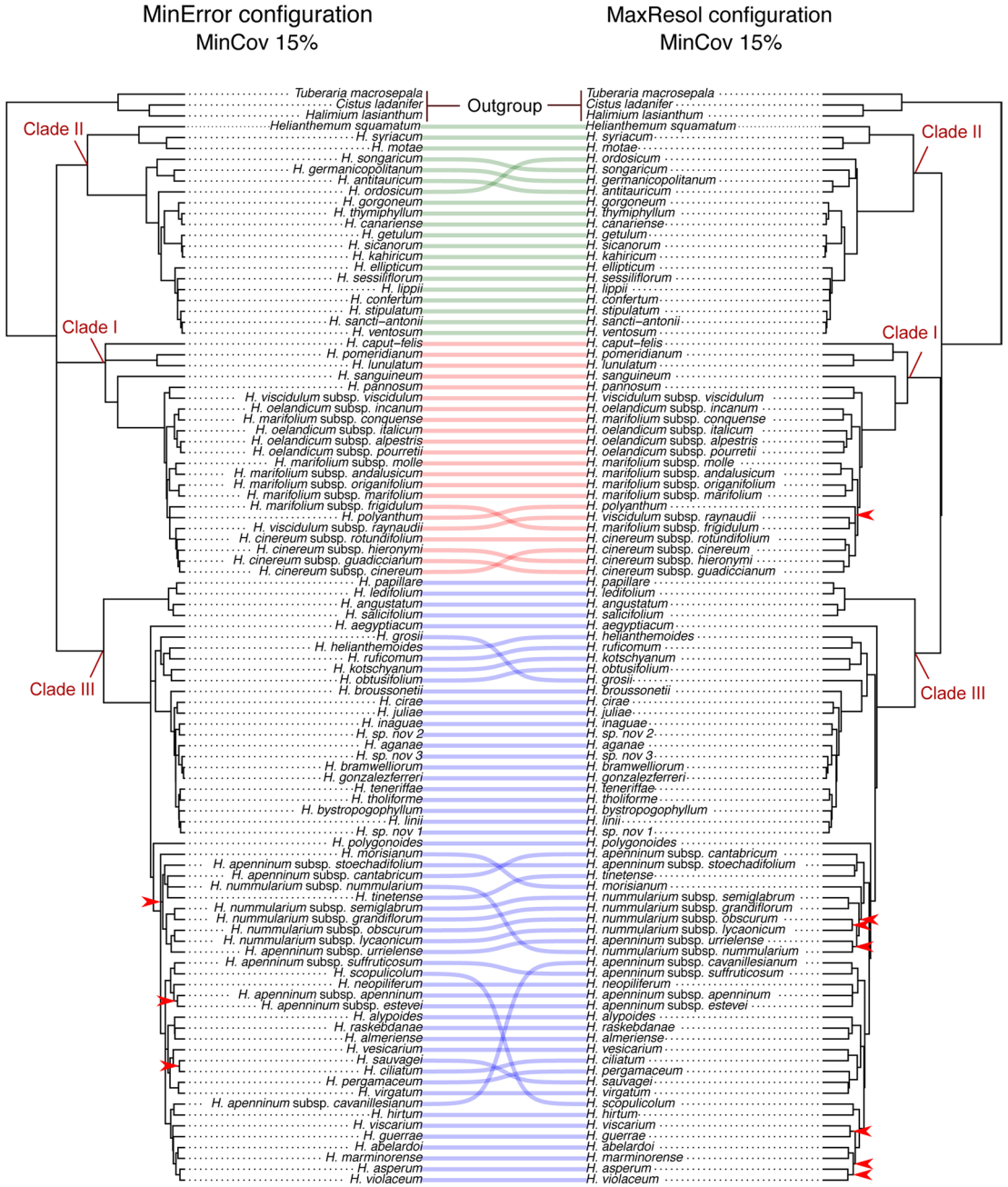

FIGURE 2 | Comparison of 50% majority-rule consensus trees resulting from Bayesian analyses of *Helianthemum* GBS data in ExaBayes using the two extreme parameter configurations (MaxResol, maximizing phylogenetic resolution; and MinError, minimizing allele and SNP error rates) under 15% minimum taxon coverage. Red arrows indicate unsupported clades (PP < 0.95). Supported incongruences between analyses are highlighted with defined coloured lines, green in Clade II, red in Clade I, and blue in Clade III. Clades I, II and III are coincident with those in Aparicio et al., 2017.

method, minimum sample coverage and clustering threshold) than by the amount of missing data (dependent on the minimum taxa coverage) (**Figure 4**; see **Methods S2** for more details regarding definition of PyRAD patameters).

The Most Robust Configuration

Even though the MaxResol configuration provided a higher degree of phylogenetic resolution than the MinError configuration under the three percentages of minimum taxon coverage (15%, 25%, and 50%; **Figure S1**, **Table 2**), MaxResol trees had high allele and SNP error rates (between four and 10 times higher than under MinError, **Table 2**), which can presumably bias terminal branch lengths (**Figure 4**). This bias would have an adverse effect on downstream analyses (**Figures S2–S4**). On the other hand, the MinError configuration under minimum taxon coverages of 25 and 50% retrieved some relationships that were

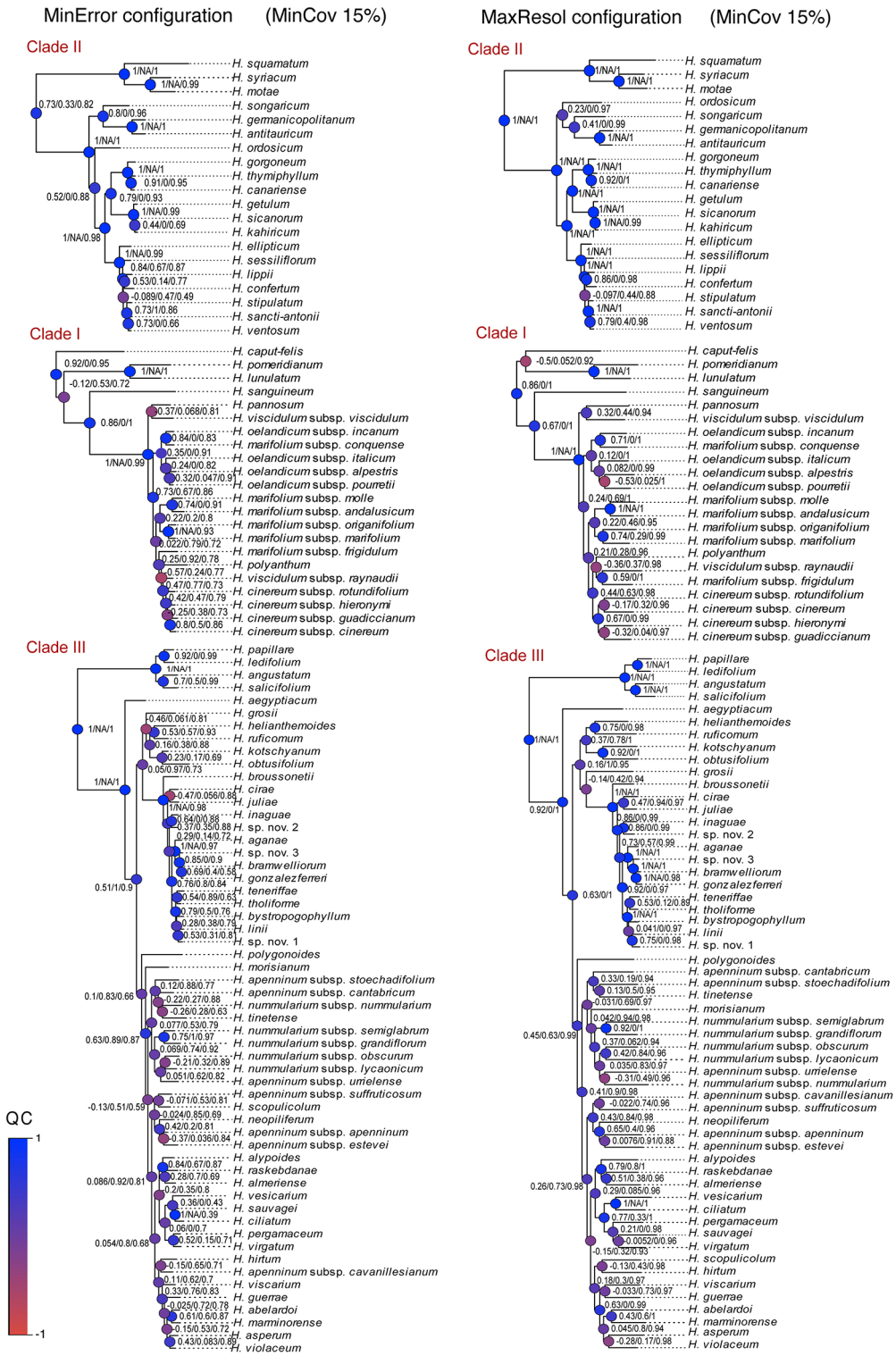

FIGURE 3 | Quartet sampling score for branches of the two Bayesian trees generated under the extreme parameter configurations (MaxResol, maximizing phylogenetic resolution; and MinError, minimizing allele and SNP error rates) under 15% minimum taxon coverage. Scores shown for each branch are in this order, QC/QD/QI. Node are coloured according to QC scores. Clades I, II and III are coincident with those in Aparicio et al., 2017.

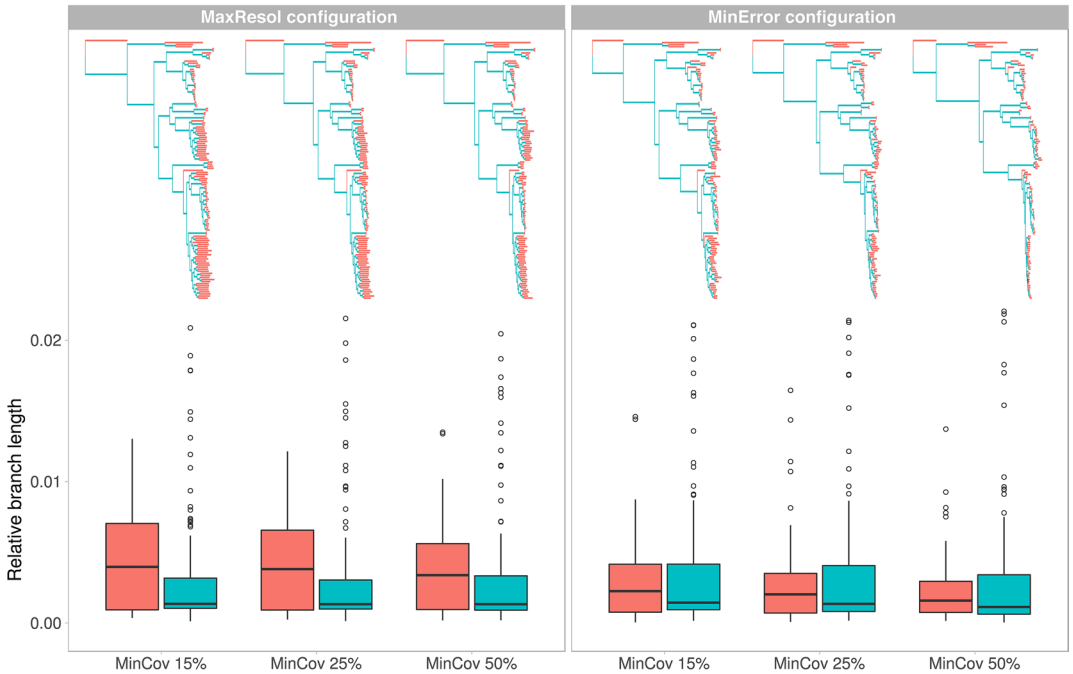

FIGURE 4 | Phylogenetic trees and variation in relative branch lengths (shown as boxplots) resulting from two extreme parameter configurations (MaxResol, maximizing phylogenetic resolution; and MinError, minimizing allele and SNP error rates) under three minimum taxon coverage percentages (MinCov 15%, 25%, and 50%) used in the assembly of GBS data of *Helianthemum*. Red represents terminal branches while blue represents internal branches.

TABLE 3 | Robinson Foulds (RF) and Branch Score (BSc) distances between Bayesian trees from MinError and MaxResol assemblies estimated in ExaBayes.

(A) Robinson Foulds (RF) distances.

		MaxResol			MinError		
		MinCov15%	**MinCov25%**	**MinCov50%**	**MinCov15%**	**MinCov25%**	**MinCov50%**
MaxResol	MinCov15%						
	MinCov25%	32					
	MinCov50%	36	20				
MinError	MinCov15%	50	56	54			
	MinCov25%	78	80	78	60		
	MinCov50%	110	110	116	100	90	

(B) Branch Score (BS) distances

		MaxResol			MinError		
		MinCov15%	**MinCov25%**	**MinCov50%**	**MinCov15%**	**MinCov25%**	**MinCov50%**
MaxResol	MinCov15%						
	MinCov25%	0.0294					
	MinCov50%	0.0569	0.0428				
MinError	MinCov15%	0.1130	0.1081	0.0716			
	MinCov25%	0.1311	0.1267	0.0892	0.0223		
	MinCov50%	0.1612	0.1582	0.1211	0.0550	0.0350	

biologically unreasonable and incongruent with those obtained from the rest of the assemblies, probably due to an extreme loss of phylogenetic signal in samples with a low starting number of reads (e.g. *H. sauvagei*, *H. kotschyanum*, *H. nummularium* subsp. *lycaonicum*; **Tables 3A, B**; **Figure S1**; **Table S4**).

Overall, we considered that the most robust species-level phylogenetic tree—taking into account degree of resolution, topological congruence with MaxResol assemblies and reliability of branch length estimation—was the phylogenetic tree resulting from the MinError configuration assembly under a minimum

taxon coverage of 15% (**Table 2**, **Figures 2**–**5**). This tree was selected as a suitable phylogenetic framework for downstream evolutionary analyses.

Phylogenetic Relationships

Despite the different degrees of phylogenetic resolution and minor incongruences obtained under the broad set of configurations and assemblies tested (**Table 2**, **Figure S1**), all the methods carried out in the present study consistently recovered similar tree topologies consisting of three main clades (I, II, and III). Interestingly, these three clades all had a similar internal structure, namely, one species-rich subclade coinciding with the larger sects. *Eriocarpum* (thereafter referred to in this paper as *Eriocarpum s.l.* in order to include its small sister section *Pseudomacularia*), *Pseudocistus* and *Helianthemum* in clades II, I, and III, respectively, accompanied by one or a few poorly diversified subclades consisting of the monospecific or species-poor sects. *Argyrolepis* and *Lavandulaceum* in clade II, *Caput-felis*, *Macularia,* and *Atlanthemum* in clade I, and *Brachypetalum* in clade III (**Figure 5**). In our reconstructions, clades II and III correspond taxonomically to subgenus *Helianthemum* and clade I to subgenus *Plectolobum*. Nomenclature and taxonomic adscriptions of taxa follow López-González (1993), but also take into account the supported systematic implications of the phylogenetic reconstruction obtained by Aparicio et al. (2017).

Downstream Analyses

Divergence Times

The extremely low values of the smoothing parameter estimated from most assemblies using TreePL ($1x10^{-199}$ to $1x10^{-9}$) indicated non-clock-like rates. All analyses recovered very narrow confidence intervals due to the low branch length variability among the 900 Bayesian trees obtained from each assembly (**Figure S2**). However, the estimated ages differed substantially between configurations and assemblies. The MaxResol configuration analysis yielded much more recent ages for the deepest nodes and older ages for shallow nodes when compared to the MinError configuration analysis (**Figures S2** and **S4**).

Diversification Rates

The overall net diversification rate of the genus *Helianthemum* (r = 0.50) was of medium magnitude, comparable to those of other Mediterranean lineages such as *Antirrhinum* (r = 0.56), *Erodium* (r = 0.20), *Genista* sect. *Spartocarpus* (r = 0.22), *Linaria* sect. *Versicolores* (r = 0.35), *Narcissus* (r = 0.17), and *Ophrys* (r = 0.55). However, net diversification rates in the three largest sections (sect. *Eriocarpum s.l.*: r = 1.11; sect. *Pseudocistus*: r = 1.26, and sect. *Helianthemum:* r = 1.61) were similar to those of some of the most rapid plant radiations in the Mediterranean Floristic Region reported to date, for example the white-flowered *Cistus* (r = 1.72), *Linaria* sect. *Supinae* (r = 1.55), the western European clade of *Erysimum* (r = 1.59), and *Reseda* sect. *Phyteuma* (r = 1.05) (see **Table 4**).

The diversification patterns estimated from BAMM analyses differed dramatically between configurations. MaxResol chronograms recovered no significant shifts in diversification rates in the tree, whilst MinError chronograms displayed very heterogeneous diversification dynamics in *Helianthemum* (**Figures 6**, **S4**). In particular, the MinError configuration produced three significant shifts to increased rates of speciation (λ) relative to background levels in the genus (λ = 0.5). The first shift was inferred at the base of sect. *Eriocarpum s.l.* (λ = 0.90; 4.20 Ma), with constant speciation over time from the stem to the present. The second and third shifts occurred at the base of sect. *Helianthemum* (λ = 0.76; 3.4 Ma) and at the base of sect. *Pseudocistus* (λ = 1.06; 2.25 Ma), characterized by exponential bursts of speciation followed by stasis or a slight drop (**Figure 6**).

DISCUSSION

Compared to the previous phylogenetic reconstruction of the genus *Helianthemum* using Sanger sequencing, in which species and subspecies were mostly recovered in polytomies (Aparicio et al., 2017), here we generated a much more robust species and subspecies-level phylogenetic tree incorporating high geographical and taxonomic representativeness, strong statistical support for taxon relationships, and accurate estimates of tree topology and branch lengths. This has been achieved following an exhaustive methodological workflow specially designed to analyse a large amount of GBS data from this recently diversified lineage. We dealt with numerous methodological challenges and concluded that minimizing error rates produces more robust phylogenetic trees than maximizing phylogenetic resolution, affecting the accuracy of downstream macroevolutionary analyses. Moreover, our phylogenetic hypothesis has important implications from both systematic and evolutionary standpoints, and provides strong support for the existence of three major lineages in *Helianthemum* that have independently radiated since the Upper Miocene in contrasting geographical and ecological contexts.

Effects of Bioinformatic Parameters on Topology and Branch Lengths

The choice of an optimal bioinformatic parameterization in phylogenomics is not straightforward due to the trade-offs between the number of loci and SNPs recovered and the error rates estimated from an assembly, especially when studying recently diversified lineages (Anderson et al., 2017). To date, most studies focussing on resolving phylogenetic relationships of recently diversified clades using GBS or RADseq data have tended to maximize the number of SNPs in order to increase the amount of phylogenetic information contained in the assembly (Wagner et al., 2013; Hou et al., 2015; Wessinger et al., 2016; Tripp et al., 2017; Lee et al., 2018). In our study, the resolution of the inferred tree topologies also increased dramatically as the data matrix increased in size, despite the concomitant increase in missing data. Thus, topologies received higher support for MaxResol configuration assemblies (both in concatenation and in coalescent methods), which contain more SNPs and PIS, than for MinError datasets (**Table 2**). Furthermore, the variation in the amount of missing data did not strongly affect tree topologies when the size of the assembly was high, particularly in the MaxResol configuration, since phylogenetic trees under

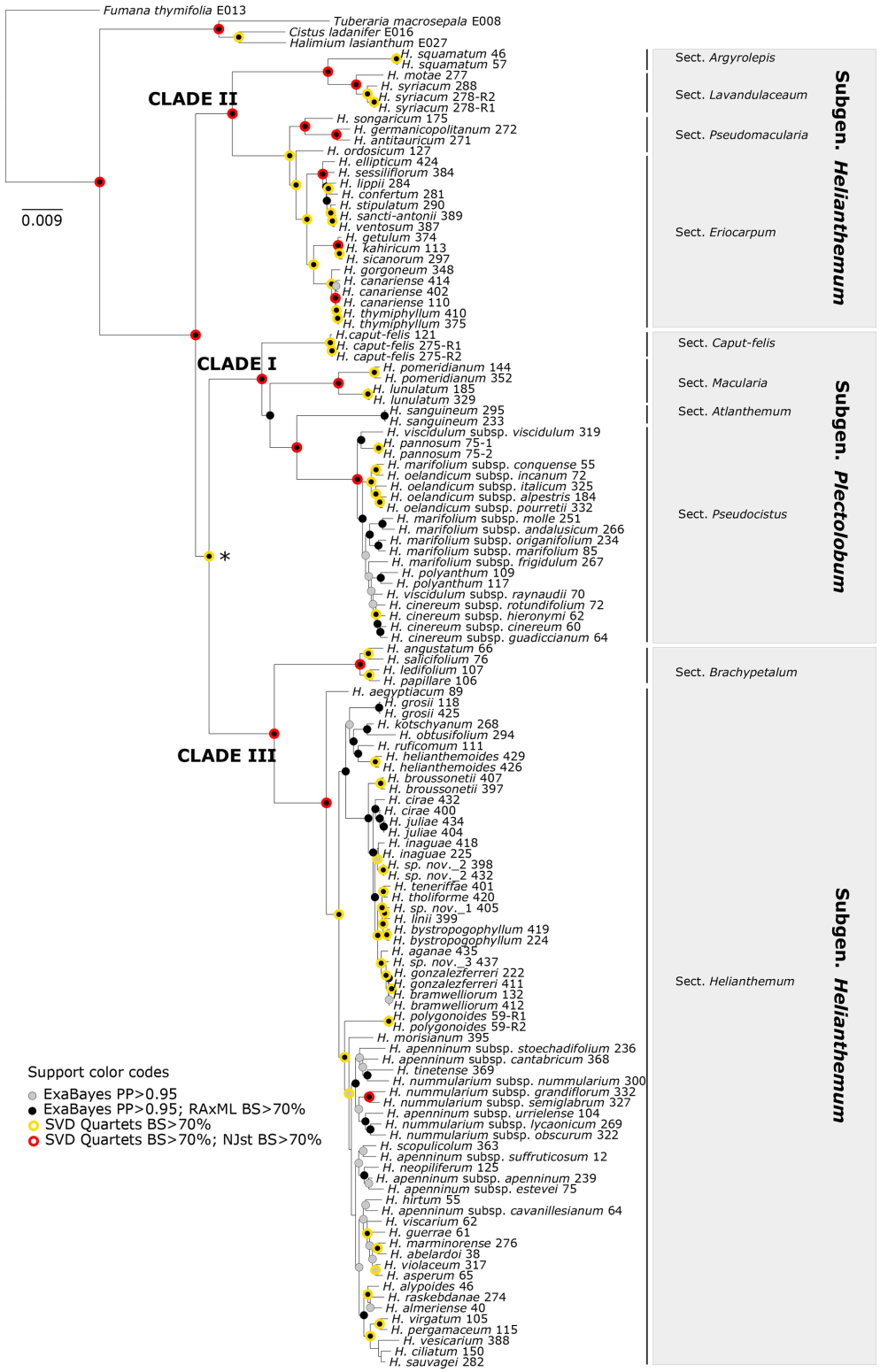

FIGURE 5 | The 50% majority-rule consensus tree obtained from Bayesian analysis of *Helianthemum* GBS data in ExaBayes using the most robust assembly (MinError configuration under 15% minimum taxon coverage). Circles of different colours indicate clades that are supported in the two concatenated (ExaBayes, RAxML) and the two coalescent (SVDquartets, NJst) phylogenetic analyses. The intrageneric taxonomic assignments of taxa (sections and subgenera) follow López-González (1993) and Aparicio et al. (2017). The asterisk denotes the single clade for which NJst provided high bootstrap support but SVDquartets did not. There were no clades with RAxML BS > 70 but ExaBayes PP > 0.95.

TABLE 4 | Diversification rates of several species-rich plant clades from the Mediterranean Basin, including the genus *Helianthemum* and its three largest sections *Eriocarpum*, *Pseudocistus,* and *Helianthemum*.

	Number of species	Crown age	Diversification rate	Distribution range	Family
Helianthemum	104	7.80 (3.56-14.08)	Medium (0.50)	Mediterranean, Macaronesia, Saharo-Arabian, Irano-Turanian	Cistaceae
Sect. *Pseudocistus*	17	1.70 (0.72–3.32)	**Fast (1.26)**	Mediterranean, Eurosiberian	
Sect. *Eriocarpum*	28	2.37 (1.01–4.63)	**Fast (1.11)**	Saharo-Arabian, Irano-Turanian, Macaronesia (Mediterranean)	
Sect. *Helianthemum*	47	1.91 (0.80–3.61)	**Fast (1.61)**	Mediterranean, Eurosiberian, Macaronesia	
Antirrhinum (Vargas et al., 2009)	20*	4.1	Medium (0.56)	W Mediterranean	Plantaginaceae
Aquilegia (European clade) (Fior et al., 2013)	25*	1.77 (0.97–2.57)	**Fast (1.47)**	S Europe	Ranunculaceae
Cistus (white-flowered) (Guzmán et al., 2009)	12	1.04 (0.06–1.41)	**Fast (1.72)**	Mediterranean	Cistaceae
Dianthus (Eurasian clade) (Valente et al., 2010)	200*	1.76 (1.09–2.43)	**Very fast (2.62)**	Mediterranean	Caryophyllaceae
Erodium (Fiz-Palacios et al., 2010)	74	18.34 (9.9–18.46)	Medium (0.20)	Mediterranean	Geraniaceae
Erysimum (W European clade) (Moazzeni et al., 2014)	25*	1.59 (0.74–2.43)	**Fast (1.59)**	W Europe	Brassicaceae
Genista sect. *Spartocarpus* (Fiz-Palacios and Valcárcel, 2013)	11	7.71 (7.18–8.23)	Medium (0.22)	C Mediterranean	Fabaceae
Linaria sect. *Supinae* (Blanco-Pastor et al., 2012)	44	2.0 (0.80–3.2)	**Fast (1.55)**	Mediterranean	Plantaginaceae
Linaria sect. *Versicolores* (Fernández-Mazuecos and Vargas, 2011)	30	7.73 (4.13–11.75)	Medium (0.35)	Mediterranean	
Narcissus (Santos-Gally et al., 2011)	70*	21.4 (16.1–27.4)	Medium (0.17)	Mediterranean	Amaryllidaceae
Ophrys (Breitkopf et al., 2015)	30*	4.9 (2.9–7.1)	Medium (0.55)	Mediterranean	Orchidaceae
Reseda sect. *Phyteuma* (Escudero et al., 2018)	16	1.98	**Fast (1.05)**	Mediterranean	Resedaceae

Number of species, crown age, diversification rates, distribution range and family are indicated for each clade. Diversification levels (slow, r < 0.1; medium, 0.1 < r < 1; fast, r > 1; see Vargas et al., 2018) are based on diversification rates calculated using Magallón and Sanderson's method based on the number of species and mean estimated crown age. Asterisks indicate uncertainty regarding species numbers. Numbers in bold represent fast or very fast diversification rates.

the three minimum taxon coverage percentages and under the two phylogenomic approaches proved to be highly congruent (**Tables 3A, B**; **Figure S1**). This result is consistent with previous observations to the effect that large amounts of missing data in reduced-representation sequencing datasets do not adversely affect the accuracy of phylogenetic inference (Rubin et al., 2012; Takahashi et al., 2014; Hou et al., 2015; Herrera and Shank, 2016; Eaton et al., 2017; Lee et al., 2018). By contrast, some incongruent relationships were retrieved among the three assemblies under the MinError configuration, with ever-decreasing biological sense as the minimum taxon coverage increased, probably due to an excessive loss of phylogenetic information from samples with a low initial number of reads (**Tables 3A, B**; **Figure S1**, **Table S4**).

Although great efforts are usually devoted to maximizing the number of SNPs in order to optimize phylogenetic resolution, the effects of error rates on phylogenetic inference are rarely explored (Clark and Whittan, 1992; Lemmon et al., 2009). NGS methods may generate twice as many sequencing errors as Sanger sequencing (Ewing and Green, 1998; Wang et al., 2012; Glenn, 2014) and reduced-representation sequencing methods are prone to a number of additional sources of error. The effects of allele and SNP errors on population genetic inferences seem to be clear, and include an inflation of nucleotide diversity and a skewing of the SNP frequency spectrum towards rare SNPs (Ho et al., 2005; Johnson and Slatkin, 2008; Pool et al., 2010). These complications can hinder a biologically meaningful interpretation of population genetic data. However, there is a lack of consensus on how error rates bias phylogenetic reconstructions, with some authors noting that confidence in a tree depends on the sequencing error rate (Clark and Whittan, 1992) and others suggesting that error rates may be less detrimental for phylogenetics than for population genetics (Anderson et al., 2017). In our study, the generally congruent topologies obtained under both parameter configurations (**Figures 2** and **3**) suggest that the differential error rates resulting from applying contrasting bioinformatic parameter values have no significant effects on phylogenetic relationships. However, datasets maximizing resolution (MaxResol) produced considerably longer terminal branch lengths compared to datasets minimizing error rates (MinError), while relative internal branch lengths remained quite constant (**Figure 4**). This could be interpreted as an artefact resulting from the fact that each tip in a MaxResol tree has extra 'substitutions' per site due to sequencing errors. In agreement with this, recent evidence indicates that sequencing errors, if not corrected, can significantly influence branch length estimates (Kuhner and McGill, 2014). Other studies have suggested that two further factors may also bias branch length estimates: the assumption of a single evolutionary model and the presence of large amounts of missing data, whose effects may be more pronounced as dataset size and complexity increase (e.g. Lemmon et al., 2009: Schwartz and Mueller, 2010; Darriba et al., 2016). Despite the fact that our study design did not permit us to discriminate whether the misestimation of branch lengths was the result of any particular factor, it is clear that maximizing phylogenetic resolution leads to higher potential bias in branch length estimation than minimizing error rates, an issue that deserves further attention.

The comparison of inferred shifts in diversification rates between MaxResol and MinError datasets (after time-calibration) revealed significantly different patterns. In particular, the

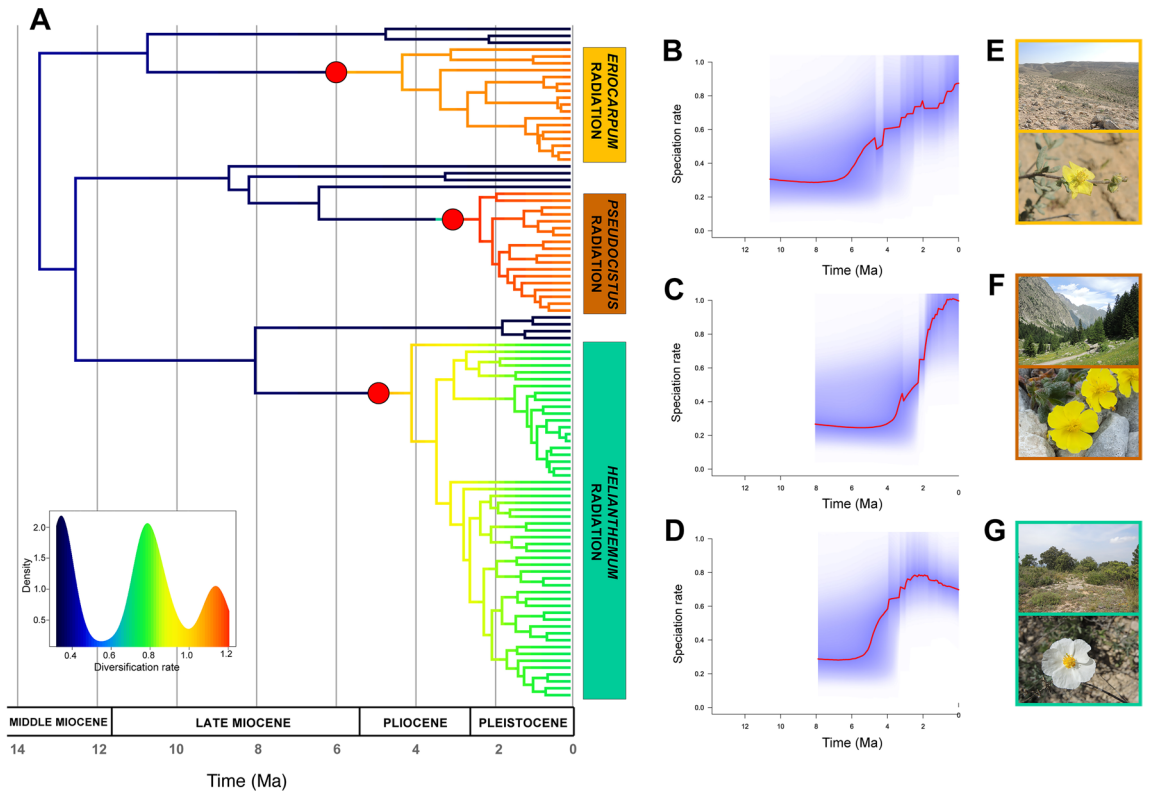

FIGURE 6 | Diversification rates in *Helianthemum* based on GBS data. **(A)** Time-calibrated phylogenetic tree obtained in TreePL from the most robust assembly (MinError configuration under 15% minimum taxon coverage), with branches coloured according to diversification rates estimated using Bayesian Analysis of The performance of the research Macroevolutionary Mixtures (BAMM). Red circles at the base of the three largest sections (*Eriocarpum s.l.*, *Pseudocistus* and *Helianthemum*) mark the three diversification rate shifts initiating three evolutionary radiations. The insert shows the density of rate values across the phylogeny. **(B–D)** Speciation rates over time estimated by BAMM for each radiation, starting from their respective stem nodes. **(E–G)** Representative species and ecosystems of the three radiations: **(E)** *Helianthemum sessiliflorum* (Desf.) Pers. in Negev Desert (Israel); **(F)** *H. oelandicum* subsp. *alpestre* (Jacq.) Ces. of alpine pastures in Alpes-Maritimes (France); **(G)** *H. apenninum* (L.) Miller subsp. *apenninum* in Mediterranean maquis at Pico Ñoño Martés (Spain).

MaxResol configuration recovered no diversification rate shifts along the tree, while the MinError configuration resulted in three accelerations of diversification rates coinciding with the origin of the three largest taxonomical sections (**Figure 6**, **Figures S3 and S4**). Thus, the artificial inflation of terminal branch lengths caused by high SNP error rates may lead to spurious interpretations of evolutionary patterns in our particular study group and probably in other clades similarly subjected to rapid diversification. Radiating lineages may be particularly susceptible to the disruption of the detection of shifts in diversification rates when biases in estimates of terminal branch lengths occur, since these lineages are characterized by short branch lengths and low pairwise sequence divergence due to closely spaced branching events (Guzmán et al., 2009; Glor, 2010). Therefore, although the topological accuracy of phylogenetic trees is important for purposes such as taxonomic classification (e.g. see discussion in de Queiroz and Gauthier, 1990), it is essential to stress that the accuracy of tree branch lengths is critical for further evolutionary inferences such as divergence time estimation, diversification rate calculation, ancestral state reconstruction, tree-dependent comparative methods and biogeographic analyses (Lemmon et al., 2009; Darriba et al., 2016).

Concatenation vs. Coalescent Approaches to GBS Phylogenetics

Researchers now routinely sequence hundreds to thousands of loci in non-model organisms using reduced-representation approaches in order to reconstruct their evolutionary histories (Giarla and Esselstyn, 2015). However, the analysis of these huge datasets involves trade-offs among computational efficiency, dataset size and simplifying assumptions (Giarla and Esselstyn, 2015) which sometimes force researchers to apply suboptimal inference methods (Kubatko and Degnan, 2007). Consequently, there is an ongoing debate among phylogeneticists as to which of the two approaches—i.e. concatenation vs. coalescent—is most appropriate for inferring phylogenies from phylogenomic datasets (Huang and Knowles, 2009; Lanier et al., 2014; Gatesy and Springer, 2014).

In our reconstructed phylogenetic trees, concatenation methods provided considerably higher phylogenetic resolution than coalescent methods for all parameter assemblies. However, they recovered high statistical support for alternative topologies resulting from a few incongruences, which mainly involved nodes in sects. *Pseudocistus* and *Helianthemum* (**Figure 2**). These results agree with previous studies in which concatenated analyses produced

anomalously high statistical support for incorrect topologies when the two most commonly used branch support methods—i.e. bootstrap (BS) and posterior probability (PP)—are applied (e.g. Jones et al., 2013; Fernández-Mazuecos et al., 2018). Spurious relationships under concatenation methods may be the result of the "fenestrated" nature of the alignment when reduced-representation data are used (i.e. high proportion of missing data; Wiens and Morrill, 2011; Roure et al., 2013; Hinchliff and Roalson, 2013) and of systematic biases (Gadagkar et al., 2005; Kumar et al., 2011). Bias may result from the specification of a single substitution model, which assumes substitution rate homogeneity across the whole dataset. Partitioned analysis may prevent this problem, but it may be computationally problematic with high numbers of loci (Fernández-Mazuecos et al., 2018). The fact that the quartet sampling analyses displayed negative QC scores for some shallow nodes (**Figure 3**) shows that this alternative branch support metric reflects topology uncertainty more accurately and is able to distinguish among different causes of incongruence between datasets (Pease et al., 2018).

Alternatively, coalescent methods produce more congruent topologies than concatenation methods, but with a generally low BS within sects. *Pseudocistus* and *Helianthemum*. Although coalescent-based methods may better reflect topological uncertainty resulting from ILS and reticulate evolution in large datasets (Anderson et al., 2017), for our dataset these methods recovered limited resolution when error rates were minimized (**Figure S1**). This lack of resolution was particularly noticeable in the trees resulting from the NJst method, which are comparable with those reconstructed using Sanger sequences (Aparicio et al., 2017). Previous studies have suggested that the short length of GBS loci (c. 100–200 bp) may result in poorly informative gene trees, which may be problematic for species tree inference (Salichos and Rokas, 2013). Although these methods may be adequate at shallow evolutionary scales (e.g. to resolve phylogenetic relationship among closely related species and populations; Fernández-Mazuecos et al., 2018), they do not seem to be suitable for establishing a robust phylogenetic framework of species-rich clades, particularly under assembly configurations that minimize error rates. In fact, software packages focused on downstream macroevolutionary analyses usually require strictly bifurcating trees (e.g. BioGeoBEARS; Matzke, 2013) which have only been recovered under concatenation methods in our study case.

Based on the topological changes (particularly at shallow nodes) that we found associated with changes in assembly parameters (i.e. clustering threshold, minimum sample coverage and minimum taxon coverage), it is still clear that conducting multiple analyses based on a range of parameter values (Takahashi et al., 2014; Leaché et al., 2015), different phylogenetic approaches and a range of branch support methods is necessary to evaluate if high clade support values provide a realistic measurement of confidence (Fernández-Mazuecos et al., 2018; Pease et al., 2018).

Systematics and Evolutionary Implications

Non-Monophyly of Taxa at Different Taxonomic Ranks

The robust phylogenetic reconstruction presented in this paper highlights the need for a comprehensive taxonomic review of the genus *Helianthemum*, from the definition of subgenera to the delimitation of species and subspecies. In particular, our study shows that the subgenus *Helianthemum* as currently defined is paraphyletic, since it is retrieved in two different non-sister clades (i.e. clades II and III). In addition, most taxonomically complex species (e.g. *H. apenninum*, *H. cinereum*, *H. marifolium*, *H. nummularium* and *H. oelandicum*), which are characterised by an array of morphological forms usually treated as subspecies (Soubani et al., 2014a, Soubani et al., 2014b; Volkova et al., 2016), are non-monophyletic (see **Figure 5**).

The topological conflicts detected for some nodes in the concatenation analyses (**Figure 2**)—particularly those involving the above-mentioned complex species—as well as the low support for the two large sects. *Pseudocistus* and *Helianthemum* in the QS and coalescent analyses (**Figures 3** and **S1**) likely reflect the fact that trait convergence, ILS, hybridization and introgression are currently playing an essential role in the differentiation of these lineages. This idea is also supported by phylogeographical approaches (Soubani et al., 2014a; Soubani et al., 2014b; Widén, 2015; Widén, 2018; Volkova et al., 2016). Future taxonomical and microevolutionary studies are therefore required to obtain more detailed insights into the processes driving species diversification and differentiation in these complex species (Martín-Hernanz et al., 2019).

Three Recent Radiating Lineages in Contrasting Geographical, Ecological and Temporal Contexts

In addition to a robust phylogenetic framework, the detection of recent evolutionary radiations requires the evaluation of the following operational criteria: 1) a recent common ancestor, 2) species-poor sister lineages, and 3) significant bursts of diversification (Nee et al., 1996; Sanderson and Donoghue, 1996; Pybus and Harvey, 2000; Schluter, 2000; Glor, 2010). Based on the first two criteria, the existence of three radiating lineages in *Helianthemum* was recently suggested by Aparicio et al. (2017). Here we provide further empirical evidence based on two analytical approaches that confirm the occurrence of significant bursts of diversification. Firstly, absolute net diversification rates calculated using the standardized method of Magallón and Sanderson (2001) reveal that diversification rates of the three largest sections of the genus *Helianthemum* (i.e. *Eriocarpum s.l.*, *Pseudocistus* and *Helianthemum*) are similar to those of other radiating lineages in the Mediterranean Floristic Region including the white-flowered clade of *Cistus* and the western European clades of *Erysimum* and *Reseda* sect. *Phyteuma* (Vargas et al., 2018; **Table 4**). Secondly, we identified three significant increases in speciation rates at the base of the above-mentioned sections (**Figure 6**).

The occurrence of multiple radiations in a large clade represents a powerful comparative system for addressing fundamental questions about patterns and processes underlying rapid diversification, as has previously been demonstrated in other plant groups (e.g. *Echium,* García-Maroto et al., 2009; *Lupinus*, Drummond et al., 2012; *Androsace*, Roquet et al., 2013). Some clues can be derived from our analysis that can help to determine whether radiations in *Helianthemum* are adaptive or not: 1) homogeneous ecological conditions in sect. *Eriocarpum s.l.* (i.e. arid and semi-arid environments

from Macaronesia, northern Africa, Horn of Africa, Anatolia, and central Asia; Aparicio et al., 2017) vs. heterogeneous in sects. *Pseudocistus* and *Helianthemum* (i.e. Mediterranean and alpine environments in Europe and western Asia; Aparicio et al., 2017); 2) Pliocene origin of sect. *Eriocarpum s.l.* vs. late Pliocene in sects. *Pseudocistus* and *Helianthemum*; and 3) constant speciation over time in sect. *Eriocarpum s.l.* vs. density-dependent cladogenesis in sects. *Pseudocistus* and *Helianthemum* (see **Figure 6**). Ongoing studies (Martín-Hernanz et al., unpublished) are specifically addressing the adaptative nature of trait evolution, biogeographic patterns and potential associations between diversification rate shifts and ancestral areas or character states on the basis of the robust phylogenetic framework here established.

DATA AVAILABILITY STATEMENT

SRA data can be found in NCBI using accession numbers in **Supplementary Table S4** or accessible with the following link (https://www.ncbi.nlm.nih.gov/sra/PRJNA573639).

AUTHOR CONTRIBUTIONS

The idea and design of the research were developed by SM-H, AA, and RGA. The performance of the research was developed by SM-H, AA, ER, and RGA. The data collection was mainly carried out by SM-H, AA, ER, AR-B, AS-G, MO-C, and RGA. The analyses and interpretation of the data were carried out by SM-H, AA, MF-M, and RGA. Finally, the manuscript was written and discussed between all authors and led by SM-H, AA, and RGA.

ACKNOWLEDGMENTS

The authors thank Ori Fragman-Sapir, Ricardo Mesa, Aurelio Acevedo, Ángel Palomares and Marco Díaz-Bertrana for their help with field sampling, and Miquel Capó Servera and Magdalena Vicens for providing plant material from the Balearic Islands. The authors also express their gratitude to the Spanish regional governments of Andalucía, Castilla–La Mancha, and Región de Murcia for granting permits to collect samples, and are especially grateful to the Canary Islands Regional Government and the following institutions for granting permits to collect samples of certain strictly protected species: Jardín Botánico Viera y Clavijo, Cabildo de Gran Canaria, Cabildo Insular de la Gomera, Cabildo Insular de La Palma, Cabildo de Lanzarote, Cabildo de Tenerife, Caldera de Taburiente National Park, and Teide National Park. Thanks are also due to Antonio Jesús Molina Jiménez and the rest of the CICA support team for providing guidance on the use of the High Performance Computing-HPC facility. Finally, the authors thank Mike Lockwood for linguistic correction. This research was funded by grants CGL2014-52459-P and CGL2017-82465-P from the Spanish Ministerio de Economía y Competitividad to AA. SM-H is currently funded by the Spanish Secretaría de Estado de Investigación, Desarrollo e Innovación (FPI fellowship, 2015). MF-M was supported by a Juan de la Cierva fellowship (Spanish Ministerio de Economía, Industria y Competitividad, reference IJCI-2015-23459).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01416/full#supplementary-material

REFERENCES

Aberer, A. J., Kobert, K., and Stamatakis, A. (2014). ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. *Mol. Biol. Evol.* 31, 2553–2556. doi: 10.1093/molbev/msu236

Anderson, B. M., Thiele, K. R., Krauss, S. L., and Barret, M. D. (2017). Genotyping-by-sequencing in a species complex of Australian hummock grasses (*Triodia*): Methodological insights and phylogenetic resolution. *PloS One* 12, e0171053. doi: 10.1371/journal.pone.0171053

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., and Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17, 81–92. doi: 10.1038/nrg.2015.28

Aparicio, A., Martín-Hernanz, S., Parejo-Farnés, C., Arroyo, J., Lavergne, S., Yesilyurt, E. B., et al. (2017). Phylogenetic reconstruction of the genus *Helianthemum* (Cistaceae) using plastid and nuclear DNA-sequences: systematic and evolutionary inferences. *Taxon* 66, 868–885. doi: 10.12705/664.5

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* 3, e3376. doi: 10.1371/journal.pone.0003376

Barry, D., and Hartigan, J. A. (1987). Statistical analysis of hominoid molecular evolution. *Stat. Sci.* 2, 191–207. doi: 10.1214/ss/1177013356

Blanco-Pastor, J. L., Vargas, P., and Pfeil, B. E. (2012). Coalescent simulations reveal hybridization and incomplete lineage sorting in Mediterranean *Linaria*. *PLoS One*. 7, e39089. doi: 10.1371/journal.pone.0039089

Bouchenak-Khelladi, Y., Renske, E., Onstein, R. E., Xing, Y., Schwery, O., and Linder, H. P. (2015). On the complexity of triggering evolutionary radiations. *New Phytol.* 207, 313–326. doi: 10.1111/nph.13331

Breitkopf, H., Onstein, R. E., Cafasso, D., Schlüter P. M., and Cozzolino S. (2015). Multiple shifts to different pollinators fuelled rapid diversification in sexually deceptive *Ophrys* orchids. *New Phytol.* 207, 377–389. doi: 10.1111/nph.13219

Chifman, J., and Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30, 3317–3324. doi: 10.1093/bioinformatics/btu530

Clark, A. G., and Whittan, T. S. (1992). Sequencing Errors and Molecular Evolutionary Analysis. *Mol. Biol. Evol.* 9, 744–752. doi: 10.1093/oxfordjournals.molbev.a040756

Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Sauné, L., Genson, G., et al. (2014). Empirical assessment of RAD sequencing for interspecific phylogeny. *Mol. Biol. Evol.* 31, 1272–1274. doi: 10.1093/molbev/msu063

Darriba, D., Weiß, M., and Stamatakis, A. (2016). Prediction of missing sequences and branch lengths in phylogenomic data. *Bioinformatics* 32, 1331–1337. doi: 10.1093/bioinformatics/btv768

de Queiroz, K., and Gauthier, J. (1990). Phylogeny as a central principle in taxonomy: phylogenetic definitions of taxon names. *Syst. Biol.* 39, 307–322. doi: 10.2307/2992353

DeFilippis, V. R., and Moore, W. S. (2000). Resolution of phylogenetic relationships among recently evolved species as a function of amount of DNA sequence: an empirical study based on woodpeckers (Aves: Picidae). *Mol. Phylogenet. Evol.* 16, 143–160. doi: 10.1006/mpev.2000.0780

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PloS Biol.* 4, 1–12. doi: 10.1371/journal.pbio.0040088

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

Drummond, C. S., Eastwood, R. J., Miotto, S. T., 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. doi: 10.1093/sysbio/syr126

Dupuis, J. R., Brunet, B. M. T., Bird, H. M., Lumley, L. M., Fagua, G., Boyle, B., et al. (2017). Genome-wide SNPs resolve phylogenetic relationships in the North American spruce budworm (*Christoneura fumiferana*) species complex. *Mol. Phylogenet. Evol.* 111, 158–168. doi: 10.1016/j.ympev.2017.04.001

Eaton, D. A. R., Spriggs, E. L., Park, B., and Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Syst. Biol.* 66, 399–412. doi: 10.1093/sysbio/syw092

Eaton, D. A. R. (2014). PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics* 30, 1844–1849. doi: 10.1093/bioinformatics/btu121

Ebel, E. R., DaCosta, J. M., Sorenson, M. D., Hill, R. I., Briscoe, A. D., Willmott, K. R., et al. (2015). Rapid diversification associated with ecological specialization in Neotropical Adelpha butterflies. *Mol. Ecol.* 24, 2392–2405. doi: 10.1111/mec.13168

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PloS One* 6, e19379. doi: 10.1371/journal.pone.0019379

Escudero, M., Balao, F., Martín-Bravo, S., Valente, L., and Valcárcel, V. (2018). Is diversification of Mediterranean Basin plant lineages coupled to karyotypic changes? *Plant Biol.* 1, 166–175. doi: 10.1111/plb.12563

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* 8, 186–194. doi: 10.1101/gr.8.3.175

Fernández-Mazuecos, M., and Vargas, P. (2011). Historical isolation versus recent long-distance connections between Europe and Africa in bifid toadflaxes (Linaria sect. Versicolores). *PLoS One* 6, e22234. doi: 10.1371/journal.pone.0022234

Fernández-Mazuecos, M., Mellers, G., Vigalondo, B., Sáez, L., Vargas, P., and Glover, B. J. (2018). Resolving recent plant radiations: power and robustness of genotyping-by-sequencing. *Syst. Biol.* 67, 250–268. doi: 10.1093/sysbio/syx062

Fior, S., Li, M., Oxelman, B., Viola, R., Hodges, S. A., Ometto, L., et al. (2013). Spatiotemporal reconstruction of the Aquilegia rapid radiation through next-generation sequencing of rapidly evolving cpDNA regions. *New Phytol.* 198, 579–592. doi: 10.1111/nph.12163

Fiz-Palacios, O., and Valcárcel, V. (2013). From Messinian crisis to Mediterranean climate: A temporal gap of diversification recovered from multiple plant phylogenies. *Perspect Plant Ecol Syst.* 15, 130–137. doi: 10.1016/j.ppees.2013.02.002

Fiz-Palacios, O., Vargas, P., Vila, R., Papadopulos, A. S. T., and Aldasoro, J. J. (2010). The uneven phylogeny and biogeography of Erodium (Geraniaceae): radiations in the Mediterranean and recent recurrent intercontinental colonization. *Ann. Bot.* 106, 871–884. doi: 10.1093/aob/mcq184

Gadagkar, S. R., Rosenberg, M. S., and Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Exp. Zool. B Mol. Dev. Evol.* 304, 64–74. doi: 10.1002/jez.b.21026

García-Maroto, F., Mañas-Fernández, A., Garrido-Cárdenas, J. A., Alonso, D. L., Guil-Guerrero, J. L., Guzmán, B., et al. (2009). Δ6-Desaturase sequence evidence for explosive Pliocene radiations within the adaptive radiation of Macaronesian *Echium* (Boraginaceae). *Mol. Phylogenet. Evol.* 52, 563–574. doi: 10.1016/j.ympev.2009.04.009

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

Giarla, T. C., and Esselstyn, J. A. (2015). The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of philippine shrews. *Syst Biol.* 64 (5), 727–740. doi: 10.1093/sysbio/syv029

Glenn, T. C. (2014). *NGS Field Guide*. URL https://www.molecularecologist.com/next-gen-fieldguide-2014/ [accessed 1 October 2018].

Glor, R. E. (2010). Phylogenetic insights on adaptive radiation. *Annu. Rev. Ecol. Evol. S.* 41, 251–270. doi: 10.1146/annurev.ecolsys.39.110707.173447

Guzmán, B., Lledó, M. D., and Vargas, P. (2009). Adaptive radiation in Mediterranean *Cistus* (Cistaceae). *PloS One* 4, e6362. doi: 10.1371/journal.pone.0006362

Harmon, L. J., Weir, J. T., Brock, C., Glor, R. E., and Challenger, W. (2008). GEIGER: investigating evolutionary radiations. *Bioinformatics* 24, 129–131. doi: 10.1093/bioinformatics/btm538

Herrera, S., and Shank, T. M. (2016). RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *Mol. Phylogenet. Evol.* 100, 70–79. doi: 10.1016/j.ympev.2016.03.010

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

Hinchliff, C. E., and Roalson., E. H. (2013). Using supermatrices for phylogenetic inquiry: an example using the sedges. *Syst. Biol.* 62, 205–219. doi: 10.1093/sysbio/sys088

Ho, S. Y. W., Phillips, M. J., Cooper, A., and Drummond, A. (2005). Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.* 22, 1561–1568. doi: 10.1093/molbev/msi145

Hou, Y., Nowak, M. D., Mirré, V., Bjorå, C. S., Brochmann, C., and Popp, M. (2015). Thousands of RAD-seq loci fully resolve the phylogeny of the highly disjunct arctic-alpine genus *Diapensia* (Diapensiaceae). *PloS One* 10, e0140175. doi: 10.1371/journal.pone.0140175

Hrynowiecka, A., and Winter, H. (2016). Paleoclimatic changes in the Holsteinian Interglacial (Middle Pleistocene) on the basis of indicator- species method – Palynological and macrofossils remains from Nowiny Zukowskie site (SE Poland). *Quatern. Int.* 409, 255–269. doi: 10.1016/j.quaint.2015.08.036

Huang, H., and Knowles, L. L. (2009). What is the danger of the anomaly zone for empirical phylogenetics? *Syst. Biol.* 58, 527–536. doi: 10.1093/sysbio/syp047

Hughes, C. E., Nyffeler, R., and Linder, H. P. (2015). Evolutionary plant radiations: where, when, why and how? *New Phytol.* 207, 249–253. doi: 10.1111/nph.13523

Jayaswal, V., Jermiin, L. S., Poladian, L., and Robinson, J. (2011). Two stationary nonhomogeneous markov models of nucleotide sequence evolution. *Syst. Biol.* 60, 74–86. doi: 10.1093/sysbio/syq076

Johnson, P. L. F., and Slatkin, M. (2008). Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.* 25, 199–206. doi: 10.1093/molbev/msm239

Jones, J. C., Fan, S., Franchini, P., Schartl, M., and Meyer, A. (2013). The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Mol. Ecol.* 22, 2986–3001. doi: 10.1111/mec.12269

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

Kuhner, M. K., and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11, 459–468. doi: 10.1093/oxfordjournals.molbev.a040126

Kuhner, M. K., and McGill, J. (2014). Correcting for sequencing error in maximum likelihood phylogeny inference. *G3.* 4, 2545–2552. doi: 10.1534/g3.114.014365

Kumar, S., Filipski, A. J., Battistuzzi, F. U., Pond, S. L. K., and Tamura, K. (2011). Statistics and truth in phylogenomics. *Mol. Biol. Evol.* 29, 457–472. doi: 10.1093/molbev/msr202

Lanier, H. C., Huang, H., and Knowles, L. L. (2014). How low can you go? the effects of mutation rate on the accuracy of species-tree estimation. *Mol. Phylogenet. Evol.* 70, 112–119. doi: 10.1016/j.ympev.2013.09.006

Leaché, A. D., Chavez, A. S., Jones, L. N., Grummer, J. A., Gottscho, A. D., and Linkem, C. W. (2015). Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biol. Evol.* 7, 706–719. doi: 10.1093/gbe/evv026

Lee, K. M., Ivanov, V., Hausmann, A., Kaila, L., Wahlberg, N., and Mutanen, M. (2018). Information dropout patterns in restriction site associated DNA phylogenomics and a comparison with multilocus Sanger data in a species-rich moth genus. *Syst. Biol.* 67, 925–939. doi: 10.1093/sysbio/syy029

Lemmon, A. R., Brown, J. M., Stanger-Hall, K., and Lemmon, E. M. (2009). The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Syst. Biol.* 58, 130–145. doi: 10.1093/sysbio/syp017

Liu, L., Wu, S., and Yu L. (2015). Coalescent methods for estimating species trees from phylogenomic data. *J. Syst. Evol.* 53, 380–390. doi: 10.1111/jse.12160

Liu, L., and Yu, L. (2010). Phybase: an R package for species tree analysis. *Bioinformatics* 26, 962–963. doi: 10.1093/bioinformatics/btq062

Liu, L., and Yu, L. (2011). Estimating species trees from unrooted gene trees. *Syst. Biol.* 60, 661–667. doi: 10.1093/sysbio/syr027

López-González, G. (1993). "*Helianthemum*," in *Flora iberica*, vol. 3. Eds. Castroviejo, S., Aedo, C., Cirujano, S., Laínz, M., Montserrat, P., Morales, R., Muñoz Garmendia, F., Navarro, C., Paiva, J., and Soriano, C. (Madrid: Real Jardín Botánico, C.S.I.C), 365–421.

Magallón, S., and Sanderson, M. J. (2001). Absolute diversification rates in angiosperm clades. *Evolution* 55, 1762–1780. doi: 10.1111/j.0014-3820.2001.tb00826.x

Mallo, D., *Multi-locus bootstrapping script*, 2015, Available in: https://github.com/adamallo/multi-locus-bootstrapping. [accessed 15 December 2017].

Mallo, D., *NJstM*, 2016, Available in: https://github.com/adamallo/NJstM. [accessed 15 December 2017].

Martín-Hernanz, S., Martínez-Sánchez, S., Albaladejo, R. G., Lorite, J., Arroyo, J., and Aparicio, A. (2019). Genetic diversity and differentiation in narrow versus widespread taxa of *Helianthemum* (Cistaceae) in a hotspot: The role of geographic range, habitat, and reproductive traits. *Ecol. Evol.* 9, 3016–3029. doi: 10.1002/ece3.4481

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., and Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Mol. Ecol. Resour.* 15, 28–41. doi: 10.1111/1755-0998.12291

Matzke, N. J. (2013). BioGeoBEARS: BioGeography with Bayesian (and likelihood) evolutionary analysis in R Scripts. *R Package version* 0.2 1, 2013. doi: 10.5281/zenodo.1478250

McVay, J. D., and Carstens, B. C. (2013). Phylogenetic model choice: justifying a species tree or concatenation analysis. *J. Phylogenetics Evol. Biol.* 1, 114. doi: 10.4172/2329-9002.1000114

Meiklejohn, K. A., Faircloth, B. C., Glenn, T. C., Kimball, R. T., and Braun, E. L. (2016). Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. *Syst. Biol.* 65, 612–627. doi: 10.1093/sysbio/syw014

Menke, B. (1976). Pliozäne und ältestquartäre Sporen- und Pollenflora von Schleswig-Holstein. *Geol. Jahrb.* 32, 3–197.

Miller, M. R., Atwood, T. S., Eames, B. F., Eberhart, J. K., Yan, Y. L., Postlethwait, J. H., et al. (2007). RAD marker microarrays enable rapid mapping of zebrafish mutations. *Genome Biol.* 8, R105. doi: 10.1186/gb-2007-8-6-r105

Moazzeni, H., Zarre, S., Pfeil, B. E., Bertrand, Y. J. K., German, D. A., Al-Shehbaz, I. A., et al. (2014). Phylogenetic perspectives on diversification and character evolution in the species-rich genus *Erysimum* (Erysimeae; Brassicaceae) based on a densely sampled ITS approach. *Bot. J. Linn. Soc.* 175, 497–522. doi: 10.1111/boj.12184

Nadeau, N. J., Martin, S. H., Kozak, K. M., Salazar, C., Dasmahapatra, K. K., Davey, J. W., et al. (2013). Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Mol. Ecol.* 22, 814–826. doi: 10.1111/j.1365-294X.2012.05730.x

Naud, G., and Suc, J. P. (1975). Contribution à l'étude paléofloristique des Coirons (Ardèche): premières analyses polliniques dans les alluvions sous-basaltiques et interbasaltiques de Mirabel (Miocène supérieur). *B. Soc Geol. Fr.* 17, 820–827. doi: 10.2113/gssgfbull.S7-XVII.5.820

Nee, S., Barraclough, T. G., and Harvey, P. H. (1996). "Temporal changes in biodiversity: detecting patterns and identifying causes," in *Biodiversity: a biology of numbers and differences*. Ed. Gaston, K. J. (Oxford, U.K: Blackwell Sciences), 230–252.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

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 (3), 385–403. doi: 10.1002/ajb2.1016

Pirie, M. (2015). Phylogenies from concatenated data: is the end nigh? *Taxon* 64, 421–423. doi: 10.12705/643.1

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS One* 7, e32253. doi: 10.1371/journal.pone.0032253

Pool, J. E., Hellmann, I., Jensen, J. D., and Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome Res.* 20, 291–300. doi: 10.1101/gr.079509.108

Pybus, O. G., and Harvey, P. H. (2000). Testing macro-evolutionary models using incomplete molecular phylogenies. *P. R. Soc B.* 267, 2267–2272. doi: 10.1098/rspb.2000.1278

Rabosky, D. L. (2014). Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PloS One* 9, 389543. doi: 10.1371/journal.pone.0089543

Rabosky, D. L., Donnellan, S. C., Grundler, M., and Lovette, I. J. (2014a). Analysis and visualization of complex macroevolutionary dynamics: an example from Australian scincid lizards. *Syst. Biol.* 63, 610–627. doi: 10.1093/sysbio/syu025

Rabosky, D. L., Grundler, M., Anderson, C., Shi, J. J., Brown, J. W., Huang, H., et al. (2014b). BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods Ecol. Evol.* 5, 701–707. doi: 10.1111/2041-210X.12199

Rabosky, D. L., Santini, F., Eastman, J., Smith, S. A., Sidlauskas, B., Chang, J., et al. (2013). Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.* 4, 1958. doi: 10.1038/ncomms2958

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

Rivers, D. M., Darwell, C. T., and Althoff, D. M. (2016). Phylogenetic analysis of RAD-seq data: examining the influence of gene genealogy conflict on analysis of concatenated data. *Cladistics* 32, 672–681. doi: 10.1111/cla.12149

Robinson, D., and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147. doi: 10.1016/0025-5564(81)90043-2

Roquet, C., Boucher, F. C., Thuiller, W., and Lavergne, S. (2013). Replicated radiations of the alpine genus *Androsace* (Primulaceae) driven by range expansion and convergent key innovations. *J. Biogeogr.* 40, 1874–1886. doi: 10.1111/jbi.12135

Roure, B., Baurain, D., and Philippe, H. (2013). Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol. Biol. Evol.* 30, 197–214. doi: 10.1093/molbev/mss208

Rowe, H., Renaut, S., and Guggisberg, A. (2011). RAD in the realm of next-generation sequencing technologies. *Mol. Ecol.* 20, 3499–3502. doi: 10.1111/j.1365-294X.2011.05197.x

Rubin, B. E. R., Ree, R. H., and Moreau, C. S. (2012). Inferring phylogenies from RAD sequence data. *PloS One* 7, e33394. doi: 10.1371/journal.pone.0033394

Salichos, L., and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497327 &ndash, 331. doi: 10.1038/nature12130

Salichos, L., Stamatakis, A., and Rokas, A. (2014). Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.* 31, 1261–1271. doi: 10.1093/molbev/msu061

Sanderson, M. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* 19, 101–109. doi: 10.1093/oxfordjournals.molbev.a003974

Sanderson, M. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302. doi: 10.1093/bioinformatics/19.2.301

Sanderson, M., and Donoghue, M. (1996). Reconstructing shifts in diversification rates on phylogenetic trees. *Trends Ecol. Evol.* 11, 15–20. doi: 10.1016/0169-5347(96)81059-7

Santos-Gally, R., Vargas, P., and Arroyo, A. (2011). Insights into Neogene Mediterranean biogeography based on phylogenetic relationships of mountain and lowland lineages of *Narcissus* (Amaryllidaceae). *J. Biogeogr.* 39, 782–798. doi: 10.1111/j.1365-2699.2011.02526.x

Santos-Guerra, A. (2014). Contribución al conocimiento del género *Helianthemum* Miller (Cistaceae) en las islas Canarias: *Helianthemum cirae* A. Santos sp. nov. y *H. linii* A. Santos sp. nov., especies nuevas para la isla de la Palma. *Vieraea* 42, 295–308.

Schliep, K. P. (2011). Phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592–593. doi: 10.1093/bioinformatics/btq706

Schluter, D. (2000). *The ecology of adaptive radiations*. Oxford, UK: Oxford University Press.

Schwartz, R. S., and Mueller, R. L. (2010). Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks. *BMC Evol. Biol.* 10, 5. doi: 10.1186/1471-2148-10-5

Seo, T. K. (2008). Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol. Biol. Evol.* 25, 960–971. doi: 10.1093/molbev/msn043

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., et al. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods Ecol. Evol.* 8, 907–917. doi: 10.1111/2041-210X.12700

Shaw, K. (2002). Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proc. Natl. Acad. Sci. U.S.A.* 99, 16122–16127. doi: 10.1073/pnas.242585899

Shi, J. J., and Rabosky, D. L. (2015). Speciation dynamics during the global radiation of extant bats. *Evolution* 69, 1528–1545. doi: 10.1111/evo.12681

Smith, S. A., and O'Meara, B. C. (2012). TreePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28, 2689–2690. doi: 10.1093/bioinformatics/bts492

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

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., et al. (2013). An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PloS One* 8, e54603. doi: 10.1371/journal.pone.0054603

Soubani, E., Hedrén, M., and Widén, B. (2014a). Phylogeography of the European rock rose *Helianthemum nummularium* (Cistaceae): Incongruent patterns of differentiation in plastid DNA and morphology. *Bot. J. Linn. Soc* 176, 311–331. doi: 10.1111/boj.12209

Soubani, E., Hedrén, M., and Widén, B. (2014b). Genetic and morphological differentiation across a contact zone between two postglacial immigration lineages of *Helianthemum nummularium* (Cistaceae) in southern Scandinavia. *Plant Syst. Evol.* 301, 1499–1508. doi: 10.1007/s00606-014-1170-1

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Sumner, J. G., Jarvis, P. D., Fernández-Sánchez, J., Kaine, B. T., Woodhams, M. D., and Holland, B. R. (2012). Is the general time-reversible model bad for molecular phylogenetics? *Syst. Biol.* 61, 1069–1074. doi: 10.1093/sysbio/sys042

Swofford, D. (2002). *PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4*. Sinauer: Sunderland, MA.

Takahashi, T., Nagata, N., and Sota, T. (2014). Application of RAD-based phylogenetics to complex relationships among variously related taxa in a species flock. *Mol. Phylogenet. Evol.* 80, 137–144. doi: 10.1016/j.ympev.2014.07.016

Tripp, E. A., Tsai, Y. H. E., Zhuang, Y., and Dexter, K. G. (2017). RADseq dataset with 90% missing data fully resolves recent radiation of *Petalidium* (Acanthaceae) in the ultra-arid deserts of Namibia. *Ecol. Evol.* 7, 7920–7936. doi: 10.1002/ece3.3274

Vachaspati, P., and Warnow, T. (2018). SVQquest: improving SVDquartets species tree estimation using exact optimization within a constrained search space. *Mol. Phylogenet. Evol.* 124, 122–136. doi: 10.1016/j.ympev.2018.03.006

Valente, L. M., Savolainen, V., and Vargas, P. (2010). Unparalleled rates of species diversification in Europe. *Proc. R. Soc. Lond. B Biol. Sci.* 277, 1489–1496. doi: 10.1098/rspb.2009.2163

Vargas, P., Carrio, E., Guzman, B., Amat, E., and Güemes, J. (2009). A geographical pattern of Antirrhinum speciation since the Pliocene based on plastid and nuclear DNA polymorphisms. *J. Biogeogr.* 36, 1297–1312. doi: 10.1111/j.1365-2699.2008.02059.x

Vargas, P., Fernández-Mazuecos, M., and Heleno, R. (2018). Phylogenetic evidence for a Miocene origin of Mediterranean lineages: species diversity, reproductive traits and geographical isolation. *Plant Biol.* 20, 157–165. doi: 10.1111/plb.12626

Volkova, P. A., Schanzer, I. A., Soubani, E., Meschersky, I. G., and Widén, B. (2016). Phylogeography of the European rock rose *Helianthemum nummularium* s.l. (Cistaceae): Western richness and eastern poverty. *Plant Syst. Evol.* 302, 781–794. doi: 10.1007/s00606-016-1299-1

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., et al. (2013). Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.* 22, 787–798. doi: 10.1111/mec.12023

Wang, X. V., Blades, N., Ding, J., Sultana, R., and Parmigiani, G. (2012). Estimation of sequencing error rates in short reads. *BMC Bioinformatics* 13, 185. doi: 10.1186/1471-2105-13-185

Wessinger, C. A., Freeman, C. C., Mort, M. E., Rausher, M. D., and Hileman, L. C. (2016). Multiplexed shotgun genotyping resolves species relationships within the North American genus *Penstemon*. *Am. J. Bot.* 103, 912–922. doi: 10.3732/ajb.1500519

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

Wiens, J. J., and Morrill., M. C. (2011). Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst. Biol.* 60, 719–731. doi: 10.1093/sysbio/syr025

Whitfield, J. B., and Kjer, K. M. (2008). Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu. Rev. Entomol.* 53, 26. doi: 10.1146/annurev.ento.53.103106.093304

Widén, B. (2015). Genetic basis of a key character in *Helianthemum nummularium*. Plant Syst. *Evol.* 301, 1851–1862. doi: 10.1007/s00606-015-1198-x

Widén, B. (2018). Inheritance of a hair character in *Helianthemum oelandicum* var. *canescens allele frequencies Natural populations. Plant Syst. Evol.* 34, 145–161. doi: 10.1007/s00606-017-1457-0

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620. doi: 10.1093/bioinformatics/btt593

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: 15 November 2019
doi: 10.3389/fpls.2019.01466

Whole Plastome Sequencing Within *Silene* Section *Psammophilae* Reveals Mainland Hybridization and Divergence With the Balearic Island Populations

José Carlos del Valle[1]*, *Inés Casimiro-Soriguer*[1], *Mª Luisa Buide*[1], *Eduardo Narbona*[1] *and Justen B. Whittall*[2]

[1] Department of Molecular Biology and Biochemical Engineering, Pablo de Olavide University, Seville, Spain, [2] Department of Biology, Santa Clara University, Santa Clara, CA, United States

OPEN ACCESS

Edited by:
Gonzalo Nieto Feliner,
Real Jardín Botánico (RJB), Spain

Reviewed by:
Eduardo Ruiz-Sanchez,
University of Guadalajara, Mexico
Mario Fernández-Mazuecos,
Real Jardín Botánico (RJB), Spain
Yamama Naciri,
Conservatoire et Jardin Botanique de la Ville de Genève, Switzerland

***Correspondence:**
José Carlos del Valle
jcdelgar@upo.es

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

Received: 03 June 2019
Accepted: 22 October 2019
Published: 15 November 2019

Citation:
del Valle JC, Casimiro-Soriguer I, Buide ML, Narbona E and Whittall JB (2019) Whole Plastome Sequencing Within Silene Section Psammophilae Reveals Mainland Hybridization and Divergence With the Balearic Island Populations. Front. Plant Sci. 10:1466. doi: 10.3389/fpls.2019.01466

Reconstructing the phylogenetic relationships within Caryophyllaceae tribe Sileneae has been obscured by hybridization and incomplete lineage sorting. *Silene* is the largest genus in the Caryophyllaceae, and unraveling its evolutionary history has been particularly challenging. In order to infer the phylogenetic relationships among the five species in *Silene* section *Psammophilae*, we have performed a genome skimming approach to acquire the complete plastid genome (cpDNA), nuclear ribosomal cistron (nrDNA), and partial mitochondrial genome (mtDNA). We have included 26 populations, representing the range of each species' distribution. This section includes five morphologically similar species endemic to the Iberian Peninsula and Balearic Islands (Ibiza and Formentera), yet some of them occupy distinct edaphic habitats (e.g. maritime sands, calcareous sandstones). In addition to phylogeographic analyses, genetic structuring using the chloroplast data set was inferred with Discriminant Analysis of Principal Components (DAPC), analyses of molecular variance (AMOVA), and a partial Mantel test. Reference-guided assembly of 50 bp single-end and 250 bp paired-end Illumina reads produced the nearly complete cpDNA genome (154 kbp), partial mtDNA genome (from 81 to 114 kbp), and the nrDNA cistron (6.4 kbp). Selected variable regions of the cpDNA and mtDNA assemblies were confirmed by Sanger sequencing. Phylogenetic analyses of the mainland populations reveal incongruence among the three genomes. None of the three data sets produced relationships consistent with taxonomy or geography. In contrast, *Silene cambessedesii,* present in the Balearic Islands, is the only species that forms a strongly supported monophyletic clade in the cpDNA genome and is strongly differentiated with respect to the remaining taxa of the Iberian Peninsula. These results contrast with those obtained for mainland populations. Across the entire analysis, only one well-supported mainland clade of *Silene littorea* and *Silene stockenii* emerges from the southern region of the Iberian Peninsula. DAPC and AMOVA results suggest the absence of genetic structure among mainland populations of *Silene* section *Psammophilae*, whereas partial Mantel test discarded spatial correlation of genetic differentiation. The widespread incongruence between morphology-based taxonomic boundaries and phylogeography suggests

a history of interspecific hybridization, in which only a substantial geographic barrier, like isolation by the Mediterranean Sea, was sufficient to create and maintain species boundaries in *Silene* section *Psammophilae*.

Keywords: allopatric speciation, Balearic Islands, genome skimming, hybridization, Iberian Peninsula, introgression, Sileneae

INTRODUCTION

The Mediterranean Basin is commonly described as one of the most important biodiversity hotspots in the world (Médail and Quézel, 1997; Médail and Quézel, 1999; Myers et al., 2000). In particular, the Iberian Peninsula and adjacent Balearic Islands emerge as key centers of biodiversity due to their complex geological history (including a great diversity of substrates such as serpentines, dolomites, and gypsum, among others) and spatially heterogeneous climate (Médail and Quézel, 1997; Thompson, 2005), making them ideal for examining biogeographic and evolutionary processes in plants. The coupling of the geographical position of the Iberian Peninsula (flanked by the Pyrenees to the north, the Atlantic Ocean to the west, and the Mediterranean Sea to the south and east) and Balearic Islands (isolated from mainland) with Mediterranean climate leads to exceptional ecological opportunity for habitat differentiation, geographic separation, and subsequent reproductive isolation (Thompson, 2005).

The Balearic Islands are especially rich in endemics, making them excellent models for understanding speciation (e.g. Juan et al., 2004). The Balearic archipelago was separated from the mainland in the Oligocene [30–25 million years ago (Ma)], yet ephemeral land bridges connecting the mainland to the islands formed during the Messinian Salinity Crisis in the Late Miocene (ca. 5.5 Ma) (Hsü et al., 1973; Krijgsman et al., 1999; Duggen et al., 2003). Due to long-term isolation after the flooding of the Mediterranean Sea, biota on these islands gradually developed into locally adapted, novel species. The scarcity of new taxa coming to the Balearic Islands after their isolation contrasts with the colonization events in other Mediterranean islands. In the Aegean islands, for instance, glaciations during the Late Pleistocene (~0.8–0.01 Ma) decreased sea levels and created land bridges that allowed the colonization of many mainland taxa (e.g. *Nigella arvensis* and *Silene gigantea* complexes, dwarf elephants, or pigmy hippopotami, among others) (Reyment, 1983; Bittkau and Comes, 2009; Du Pasquier et al., 2017), whereas Balearic Islands remained isolated because no land bridges connected them to the mainland during this time (Van der Geer et al., 2010).

In the Iberian Peninsula, dramatic geological and climatic changes (e.g. Quaternary glaciations) have repeatedly caused fragmentation, contraction, and expansion of species ranges. In this context, recently diverged lineages may have experienced secondary contact and increased chances for hybridization and introgression (Thompson, 2005; Nieto Feliner, 2014). Hybridization is a prominent force in plant evolution that allows them to acquire genetic novelties faster than through mutations alone, creating opportunities for adaptive evolution (Arnold, 1997; Rieseberg, 1997; Rieseberg et al., 2003; Mallet, 2005). Introgressive hybridization, whether it be adaptive or neutral, may distort phylogenetic relationships in plant species where reproductive isolation is incomplete. In plants, molecular studies have tried to overcome this issue by using chloroplast sequences in addition to information from nuclear DNA sequences. Hence, incongruences between nuclear- and chloroplast-based phylogenetic trees are frequently interpreted to be the result of introgressive hybridization (e.g. Soltis and Kuzoff, 1995; Okuyama et al., 2005; Wang et al., 2016). However, phylogenetic incongruences can also be caused by incomplete lineage sorting, especially when the speciation is rapid, recent, and without persistent bottlenecks (Frajman et al., 2009a). Disentangling hybridization and incomplete lineage sorting long inhibited interpretation of incongruent molecular data sets, yet this is a matter of ongoing research, and several methods have been developed in recent years to differentiate these two historical processes (e.g. Holland et al., 2008; Joly et al., 2009).

In the tribe Sileneae (Caryophyllaceae), ancient and recent hybridization events have been proposed to be important processes that must be considered when inferring phylogenetic relationships (Erixon and Oxelman, 2008). Several studies have stressed the importance of hybridization to understand the evolutionary history of many groups within Sileneae (e.g. Erixon and Oxelman, 2008; Rautenberg et al., 2010; Petri and Oxelman, 2011; Petri et al., 2013). The ability of *Silene latifolia* and *Silene dioica* to hybridize is one of the best examples of incomplete reproductive isolation in this group (Bernasconi et al., 2009). These two closely related species show strong differences in their morphology and ecological preferences (e.g. *S. latifolia* has white flowers and grows in dry and disturbed habitats, whereas *S. dioica* has red flowers and inhabits moister soils), but these two lineages hybridize when in sympatry (Baker, 1948). Thus, their ecological and morphological differences suggest they are unique lineages, yet the lack of genetic differentiation in sympatry is the one hallmark of introgressive hybridization (Minder et al., 2007; Hathaway et al., 2009).

Inferring phylogenetic relationships within Sileneae is complicated, and the resulting phylogenies are frequently incongruent with morphology-based classifications (Oxelman and Lidén, 1995; Oxelman et al., 1997; Oxelman et al., 2001; Frajman et al., 2009b; Naciri et al., 2017). The tribe Sileneae is subdivided into eight genera (*Agrostemma* L., *Atocion* Adans., *Eudianthe* (Rchb.) Rchb., *Heliosperma* Rchb., *Lychnis* L, *Petrocoptis* A. Braun ex Endl., *Silene* L., and *Viscaria* Bernh.) (Oxelman et al., 2001), of which the genus *Silene* is the most diverse, with approximately 470 species (Oxelman et al., 2013; Petri et al., 2013), although some studies suggest up to 700 species (e.g. Greuter, 1995).

Several phylogenetic studies subdivided this genus into two well-supported subgenera: subgenus *Silene* and subgenus *Behenantha* (Popp and Oxelman, 2004; Popp and Oxelman, 2007; Erixon and Oxelman, 2008; Frajman et al., 2009a; Rautenberg et al., 2010; Rautenberg et al., 2012). However, these analyses based on chloroplast loci, nuclear ribosomal regions, and low-copy nuclear DNA led to unresolved phylogenetic relationships within each subgenus, probably due to ancient and recent hybridization (e.g. Frajman and Oxelman, 2007; Erixon and Oxelman, 2008; Rautenberg et al., 2010). Hybridization has been documented in several groups within *Silene*, for instance, in *Silene* section *Physolychnis* (Petri and Oxelman, 2011), Section *Melandrium* (Rautenberg et al., 2010), and Section *Otites* (Balounova et al., 2019), as well as in polyploid *Silene* from North America (Popp and Oxelman, 2007). Yet, the more remarkable event is the introgression between species from distinct subgenera about 6.6 Ma after the divergence of the two subgenera had occurred (Petri et al., 2013).

Silene section *Psammophilae* (Talavera) Greuter is a monophyletic group within the subgenus *Behenantha* that is composed of five species endemic to the Iberian Peninsula and Balearic Islands: *Silene adscendens* Lag., *Silene cambessedesii* Boiss. & Reut., *Silene littorea* Brot., *Silene stockenii* Chater, and *S. psammitis* Link ex Spreng. (Oxelman et al., 2013) (**Figure 1**). This section was previously considered a subsection within section *Erectorefractae* Chowdhuri (Talavera, 1979). However, Greuter (1995) proposed the sectional status for subsection *Psammophilae* based on differences in cell shape and flowering time with respect to other members of the section *Erectorefractae*, in addition to the differences previously described by Talavera et al. (1979) (e.g. monochasium inflorescences in *Psammophilae* and dichasium in *Erectorefractae*). *Silene pendula* L. was also previously included in *Silene* section *Psammophilae*, but it was placed in section *Behenantha* Otth by Oxelman et al. (2013). In addition, analysis based on ITS supports the monophyly of *S. adscendens*, *S. cambessedesii*, *S. littorea*, *S. psammitis*, and *S. stockenii* (Casimiro-Soriguer, 2015).

The basic chromosome number of species of the Section *Psammophilae* is n = 12, and the two species with available information are diploids, 2n = 24 (Talavera, 1990). They are self-compatible and are mainly pollinated by insects, although low levels of autonomous self-pollination may exist. They are all annual species, glandular-pubescent, with the inflorescence consisting of a monochasial cyme; but they differ in seed-coat ornamentation and in the length of the calyx and carpophore in fruit (Talavera, 1990; Casimiro-Soriguer, 2015). In addition to differences in their phenotypic traits, these five species have non-overlapping geographical distributions and distinct edaphic affinities. *S. littorea* grows in coastal sandy substrates along a coastal fringe from the northwestern to southeastern regions of the Iberian Peninsula. *S. cambessedesii* occurs in the same habitat on the Mediterranean islands of Ibiza and Formentera (the two largest western islands of the Balearic Islands), but it is also known from a few populations on the east coast of the

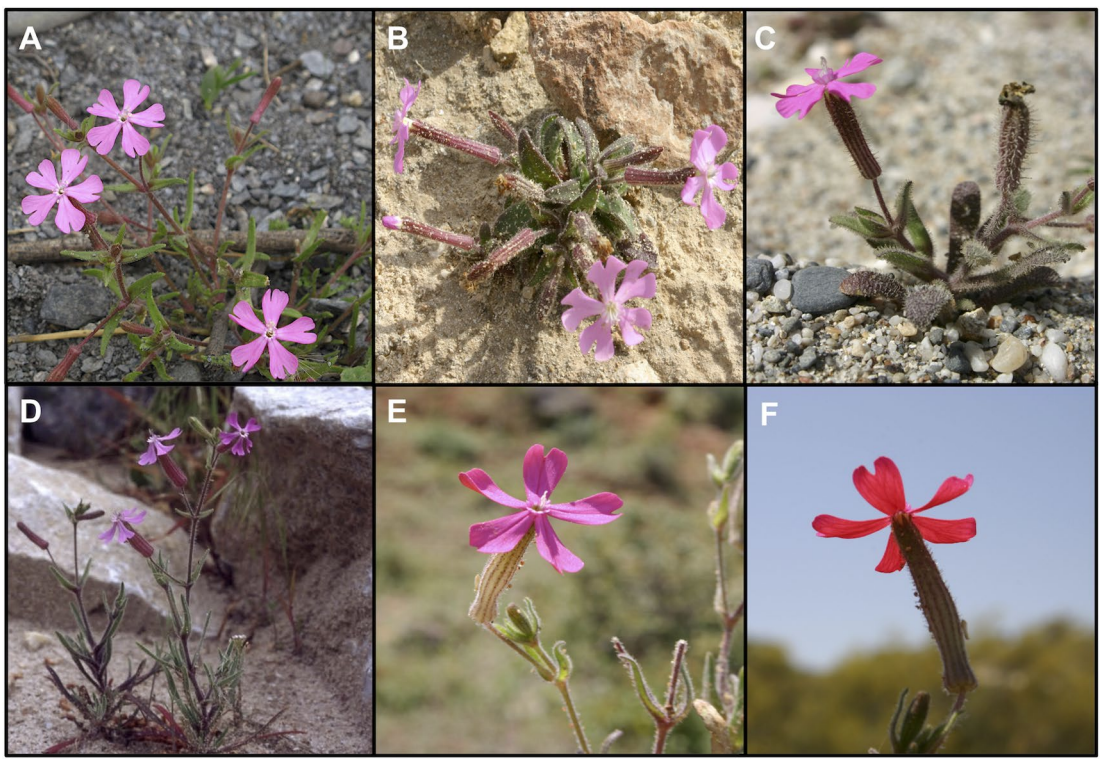

FIGURE 1 | Photographs of the five species belonging to *Silene* section *Psammophilae*. **(A)** *Silene adscendens*, **(B)** *Silene cambessedesii*, **(C)** *Silene littorea*, **(D)** *Silene psammitis*, and **(E, F)** *Silene stockenii* (showing the characteristic upper and lower side of petals).

Iberian Peninsula (Almenara, Comunidad Valenciana). *S. adscendens* occurs on intermittent streams of the southeastern Iberian Peninsula. *S. stockenii* grows in calcareous sandstones and is an endangered species restricted to the southern end of the Iberian Peninsula. Finally, *S. psammitis* is distributed throughout the Iberian Peninsula on granite or slates (subsp. *psammitis*) or dolomitic sands, and rarely on clay and serpentine (subsp. *lasiostyla*) between 300 and 1,500 m (**Figure 2**) (Talavera, 1979).

Here, we aim to unravel the phylogenetic relationships of species in *Silene* section *Psammophilae* using next-generation DNA sequencing. Recent improvements in DNA sequencing have made it possible to sequence nearly complete organellar genomes from genomic DNA (genome skimming), which has helped to infer organismal phylogenies at low taxonomic levels across different groups of angiosperms (e.g. Parks et al., 2009; Whittall et al., 2010; Kane et al., 2012; Ruhsam et al., 2015). In this study, we will address the following questions: Can genome skimming resolve the relationships of mainland and island species of *Silene* section *Psammophilae*? Do the relationships align with morphology-based species boundaries, or do they reflect geographic distance? Is genetic differentiation among these lineages correlated with geography (specifically between mainland and island populations)? Can biogeographic events (i.e. Messinian Salinity Crisis) explain the colonization of Balearic Islands by the island species of this section based on the estimating timing of mainland–island divergence? We employ sequence data from the complete plastome (cpDNA), nuclear ribosomal cistron (nrDNA), and partial mitochondrial genome (mtDNA) to address these questions using phylogenetic analysis and population genetics.

MATERIALS AND METHODS

Sampling, Genomic DNA Extraction, and Sequencing

Fresh leaf samples were collected in 2010–2012 from a total of 26 natural populations spanning the geographical range of species in *Silene* section *Psammophilae* (**Figure 2**). For each population, the DNA of five individuals was extracted using DNeasy Plant Mini Kit (Qiagen, Valencia, USA) and pooled into a single sample. DNA concentration and purity were measured with a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Inc., Wilmington, USA). Total genomic DNA was used to prepare next-generation sequencing libraries following the Nextera kit (Illumina, San Diego, USA), barcoded with 6 (single-end reads) and 8 (paired-end reads) bp indices, and sequenced at the Epigenome Center of the University of Southern California. Four libraries were pooled in equimolar concentrations and shared a half lane of an Illumina HiSeq 2000 in which they were sequenced with 50 bp single-end reads. Fortuitously, these four libraries were sequenced twice. The remaining 22 libraries were indexed

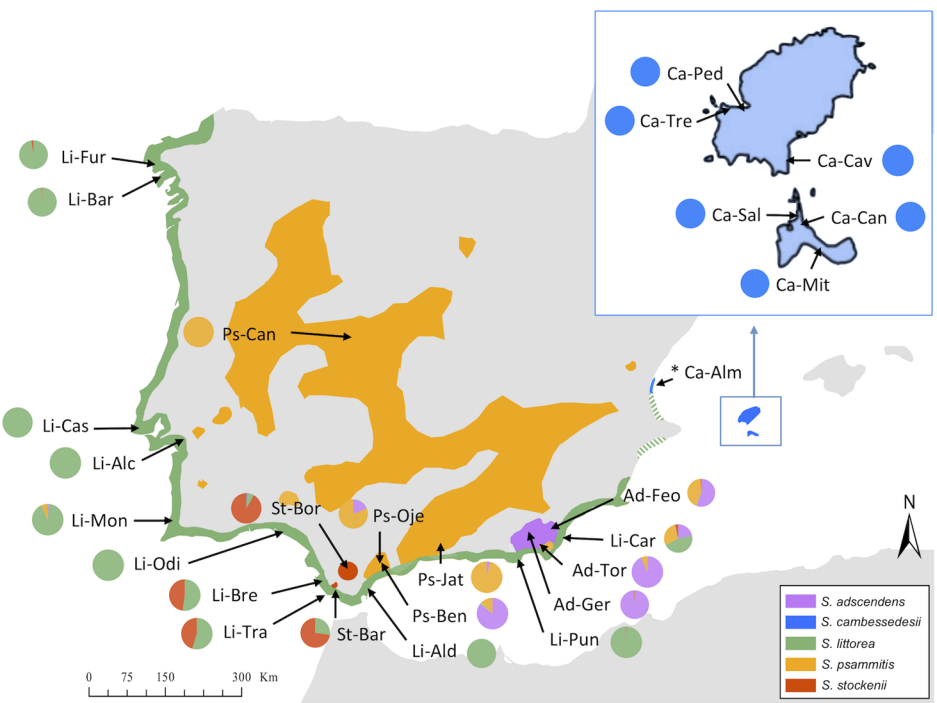

FIGURE 2 | Geographical distribution of studied populations of *Silene* section *Psammophilae* in the Iberian Peninsula and Balearic Islands and Discriminant Analysis of Principal Components (DAPC) results. Colored areas represent the distribution area of each species. The colored pie graphs represent the membership probability of each species according to the DAPC analysis on the complete plastid genome (cpDNA) when using species categories as priors. Species and population codes are shown in **Table 1**.

and then sequenced on a single lane on an Illumina HiSeq 2500 (San Diego, USA), which produced 250 bp paired-end reads.

Reference-Guided Assemblies (cpDNA, mtDNA, and nrDNA)

Prior to assembly, 3' and 5' ends of sequences with more than a 5% chance of an error per base were trimmed in Geneious v.8.1.6 (Biomatters Ltd., Auckland, New Zealand) to remove low-quality regions. Then, we conducted a reference-guided assembly using the cpDNA and mtDNA genomes of *S. latifolia* as the reference (GenBank accession numbers NC_016730 and NC_014487, respectively) and a chimeric *Silene* nrDNA cistron (see details below). We used the Geneious assembler under default settings with medium-low sensitivity and 10 iterations (Kearse et al., 2012). Regions with less than 5× sequence coverage were considered as missing data. A consensus sequence was generated for each population using an 85% consensus threshold. Thus, ambiguity codes were applied for sites below 85% consensus arising from sequencing errors or due to a variable site in the pool of five individuals per population, assuming approximately equal sequencing coverage among the five individuals pooled per sample. Annotations were transferred using a 75% similarity cutoff to the reference genome. Sequences were aligned using the MAFFT (Katoh and Standley, 2013) plugin within Geneious with default settings. Finally, regions of questionable alignment were manually adjusted or masked before subsequent analyses.

No complete nrDNA cistron sequence was available as a reference from any single species of *Silene*. Thus, we created a chimeric reference sequence following a similar procedure described in Ripma et al. (2014). We downloaded from GenBank the complete 18S (1,733 bp; AF207027) and 28S (3,332 bp; AF479084) from *Stellaria media*, the closest relative with complete sequences for these regions. We combined these with the 5.8S and both internal transcribed spacer regions (ITS1 and ITS2) from *S. littorea* (832 bp; FN821094). In addition, we performed Sanger sequencing to obtain the complete 5.8S gene with both internal transcribed spacer regions from a subsample of *Silene* section *Psammophilae* (Casimiro-Soriguer, 2015). All aforementioned nrDNA sequences were aligned, and the consensus sequence was extracted following the same settings previously described for the cpDNA. The resulting sequence was used as the nrDNA reference during the reference-guided assembly process following the parameters described above.

Single Nucleotide Polymorphism Validation

Sanger sequencing was performed in order to validate putative SNPs (single nucleotide polymorphism) that were discovered after aligning the next-generation sequencing data. Since each population was represented by a pool of five individuals, ambiguities could represent genetic variation within the pool. Thus, we individually amplified and sequenced as many individuals as possible from several of the population pools. Within each genome, we selected a specific region with the highest number of phylogenetically informative sites. For the cpDNA, we designed primers specific to the *trnK* region to amplify and sequence an 814 bp fragment that spanned 60 putative SNPs across 26 samples (*trnK-F*: GCTCGTTGCTTATTCTTTCCACA and *trnK-R*: ACTTTTGTTGGATTGGCGCT). For the mtDNA, primers were designed for the *atp1* region in order to amplify a 753 bp fragment with 125 putative SNPs (*atp1-F*: GAGTCGCAGCATCAAGGTCT and *atp1-R*: GCGGTAGATAGCCTGGTTCC). PCR conditions followed those of Dick et al. (2011) using *Taq* polymerase (New England Biolabs, Ipswich, USA) with the following thermal cycling steps: initial denature at 95°C for 3 min; 35 cycles of 95°C for 30 s, 50°C (*trnK*) and 67°C (*atp1*) for 30 s, 72°C for 90 s, a final extension at 72°C for 10 min, and a 4°C hold. PCR products were purified using exoSAP (Thermo Fisher, Cleveland, USA) and sequenced using Big Dye Terminator methodology on an ABI 3730xl DNA Analyzer (Sequetech Corp., Mountain View, USA). Single contigs were created by aligning forward and reverse reads. Contigs were then aligned to the next-generation sequences in Geneious to determine the validity of the putative SNPs.

Phylogenetic Analyses

Phylogenetic reconstruction was performed using both maximum likelihood (ML) and Bayesian approaches in RAxML (Stamatakis, 2014) and MrBayes, respectively (Huelsenbeck and Ronquist, 2001). For the ML analysis, we used the GTR+CAT approximation of the GTR+G model of nucleotide evolution with estimate of proportion of invariable sites and 1,000 bootstrap replicates. For the Bayesian analysis, we applied the GTR+G+I model of nucleotide evolution for two separate runs, each consisting of four independent chains run for 10,000,000 generations sampling every 50,000 generations after 1,000,000 generations of burn-in. Bayesian analysis runs were checked for proper mixing and convergence using Tracer v.1.6 (Rambaut et al., 2014). The cpDNA-based tree was rooted using the complete genomes of *S. latifolia* and *Silene vulgaris* (GenBank accession numbers NC_016730 and NC_016727). The low sequencing depths and subsequent assembly challenges for the mtDNA genome limited the number of sites that could be unambiguously aligned (see *Results*). Therefore, we selected six mitochondrial coding regions representing a range of substitution rates previously used in *Silene* (Barr et al., 2007; Sloan et al., 2008; Sloan et al., 2009; Rautenberg et al., 2012): the protein-encoding ATP synthase subunit 1 (*atp1*), ATP synthase subunit 4 (*atp4*), ATP synthase subunit 6 (*atp6*), cytochrome b (*cob*), cytochrome c oxidase subunit 3 (*cox3*), and NADH dehydrogenase subunit 9 (*nad9*). The mtDNA-based tree was rooted using a concatenation of these six mitochondrial genes from *S. latifolia* (extracted from NC_014487) and *S. vulgaris* (extracted from chromosomes 1 and 2 of mtDNA genome; JF750427 and JF750429, respectively). The nrDNA-based tree was rooted using the ITS regions of *S. latifolia* and *S. vulgaris* (FJ384022 and KJ918500, respectively). The resulting trees were visualized using FigTree v.1.4.2 (Rambaut, 2014). Bootstrap values (BSs) ≥ 85/posterior probabilities (PPs) ≥ 0.90 were considered as strong support, while values of 70–85% BS and 0.80–0.90 PP were considered as moderate support. In addition to the ML and Bayesian analyses, we explored phylogenetic uncertainly in the cpDNA with a NeighborNet

network in SplitsTree 4 v. 4.13.1 (Huson and Bryant, 2006) with uncorrected P distances and ambiguous sites treated with the "Average States" option.

Gene Tree–Species Tree Reconciliation

A species tree was estimated in *BEAST (Heled and Drummond, 2010) implemented in BEAST v.2.4.2 (Bouckaert et al., 2014). One limitation of *BEAST is the coalescence requirements based on which you should assign each sample to one of the five morphologically defined species. Although this could produce meaningless results if the morphologically defined species do not exist, we are confident that the morphological and ecological uniqueness of these lineages justify such *a priori* assignments. We used two partitions of the cpDNA genome with linked genealogies: the third codon position and non-coding regions of cpDNA, and the first and second codon position of the coding regions of the cpDNA. MCMC analysis was run for 500 million generations, sampling every 50,000 generations, using Yule speciation tree prior and the most appropriate nucleotide substitution model for each partition as chosen by jModelTest v. 2.1.10 (Darriba et al., 2012) under the Akaike information criterion (AIC). The selected nucleotide substitution models were GTR+G+I for both the third codon position and non-coding regions of the cpDNA and for the first and second codon position of the coding regions of the cpDNA. We used an uncorrelated log-normal relaxed molecular clock and the estimated separation of *S. latifolia* and *S. vulgaris* as 5.36 Ma (Frajman et al., 2009a) with a normally distributed standard deviation of 1.0, as calibration. The first 10% of trees were used as burn-in. Convergence and mixing were assessed in Tracer v.1.6 (Rambaut et al., 2014), with all ESS values above 120. Trees were summarized in a maximum clade credibility tree using TreeAnotator v.2.4.2 (Drummond et al., 2012). Species trees were visualized in DensiTree v.2.2.5 (Bouckaert and Heled, 2014).

Population Genetic Structure

Population structuring within the *Silene* section *Psammophilae* was explored using variable sites of the whole cpDNA genome. Since the cpDNA data set had the most variable sites (>24×) and the strongest signal (cpDNA phylogeny has more than four times the number of nodes with strong support compared to mtDNA or nrDNA), we have only analyzed the cpDNA at the population level. We used the Discriminant Analysis of Principal Components (DAPC) (Jombart, 2008; Jombart et al., 2010) to study population subdivision. We ran two clustering analyses to assess introgression among populations using both the species categories and locations as prior categories. We used species categories as priors to test the likelihood of correct assignment of populations to each species, whereas using locations as priors, we tested whether populations cluster by their geographical proximity. The number of principal components was set according to an alpha-score optimization (i.e. trade-off between power of discrimination and overfitting) (Jombart and Collins, 2015). DAPC analyses were performed using the "adegenet" package v.2.0.0 (Jombart, 2008) for the R software v.3.2.3 (R Core Team, 2016). Population structuring was explored using the "clustering with linked loci" option implemented in BAPS v.6.0 (Corander et al., 2008), which unlike STRUCTURE allows for linked loci. The number of genetically homogeneous groups was estimated from the PP (log marginal likelihood of the best partition) for three iterations of $K = 1–26$ (the total number of populations).

We tested for a correlation between genetic variation in the cpDNA genome and geographic distance between populations using a partial Mantel test (Mantel, 1967). The analysis was performed in R software v.3.2.3 (R Core Team, 2016) using the "vegan" package v.2.5.2 (Oksanen et al., 2018), with 100,000 permutations to test for significance. Tests were performed on all populations, as well as separately on just the mainland and just the island populations to dissect the relative contributions of these two groups on any isolation-by-distance findings. Additionally, analyses of molecular variance (AMOVA) (Excoffier et al., 1992) were conducted using Arlequin v.3.5 (Excoffier and Lischer, 2010). An AMOVA was conducted to assess genetic differentiation in the cpDNA genome among all studied species. A second AMOVA with just mainland species was performed to exclude the influence of island populations in the analysis. Finally, a third AMOVA was carried out to study genetic differentiation among islands. *F*-statistics (Wright, 1951) were used to estimate the proportion of genetic differentiation found among species, with significant levels determined by 1,023 permutations.

Similarity of *S. cambessedesii* Plants From Mainland and Balearic Populations

We assessed genetic similarities of *S. cambessedesii* from Balearic Islands with plants from a mainland population from the Iberian Peninsula (Almenara; **Figure 2**). Almenara is the only remaining natural population of *S. cambessedesii* on the mainland; however, plants from this population were not collected during the initial sampling for next-generation sequencing given their endangered status in the region (Comunidad Valenciana government) (Navarro et al., 2015). DNA was obtained from 20 seeds (provided by the Servicio de Vida Silvestre-CIEF in Valencia, Spain) of botanical garden grown plants from this population. Since we received seeds after the preparation of the next-generation sequencing libraries, we amplified and sequenced only four regions representing the three genomes. From the chloroplast genome, we amplified the *trnK* (see previous section *Single Nucleotide Polymorphism Validation*) and *ycf1* (871 bp) regions (using the primers *ycf1-F*: CAGTTTTTCCATTGAGTCCGTCC and *ycf1-R*: TCCCGAAAACGACCCCATTT). From the mitochondrial genome, we amplified and sequenced a fragment of the *atp1* gene (see previous section *Single Nucleotide Polymorphism Validation*). From the nuclear genome, we amplified and sequenced the ITS region (using the primers *ITS5** and *ITS26S-25R* as in Whittall et al., 2006). PCR conditions and purification were the same previously described, but using 55°C and 59°C annealing temperature for ITS and *ycf1* regions, respectively.

Finally, we examined the phylogenetic relationships between mainland and island individuals of *S. cambessedesii* by building

ML trees in RAxML for each fragment of the cpDNA (*trnK* and *ycf1*), mtDNA (*atp1*), and nrDNA (ITS) genomes. ML analyses were performed using the GTR+CAT approximation of the GTR+G model of nucleotide evolution with estimate of proportion of invariable sites and 1,000 bootstrap replicates. ML trees were rooted using *S. latifolia* and *S. vulgaris* as outgroups. The *trnK*, *ycf1*, and *atp1* fragments were extracted from the complete chloroplast and mitochondrial genomes of *S. latifolia* (NC_016730 and NC_014487 for cpDNA and mtDNA genomes, respectively) and *S. vulgaris* (NC_016727 and JF750427 for cpDNA and mtDNA genomes, respectively), whereas the ITS regions of *S. latifolia* and *S. vulgaris* were obtained from GenBank (FJ384022 and KJ918500, respectively).

RESULTS

Genomic DNA Extraction and Sequencing

Single-end sequencing generated two data sets that were merged since similar results were attained from each individual data set, obtaining an average of 25.91 million reads per sample (range 24.3 million—28.0 million; **Supplementary Table S1**). Of those, approximately 1.70 million raw reads (6.6%) were trimmed prior the assemblies. Paired-end sequencing provided an average of 1.47 million reads per sample (range 0.79 million–2.22 million; **Supplementary Table S1**). An average of approximately 60,000 raw reads were trimmed from each sample (4.0%). Raw data are available from GenBank's Short Read Archive (accession number PRJNA558348). GenBank accession numbers for cpDNA, mtDNA, and nrDNA genomes are available in **Supplementary Table S2**.

Reference-Guided Assemblies (cpDNA, mtDNA, and nrDNA)

For the cpDNA, a nearly complete chloroplast assembly from each population was recovered (alignment length = 154,199 bp). Due to alignment ambiguities, we masked ∼1% of the total length (average = 1,535.6 bp; range 1,444–1,606 bp). Total cpDNA sequencing depths were between 49.9X and 2,064.8X, with a mean of 359.91X (median = 153.95X; **Supplementary Table S1**), and there were 6,322 variable sites not including the outgroup samples. For the mtDNA genome, a mitochondrial sequence from each sample was recovered (alignment length = 254,270 bp). However, on average, only 38.0% of the alignment was assembled for each individual sample (range 31.8–44.9%), mainly corresponding to coding regions. Sequencing depths were lower than for the cpDNA, ranging from 3.8X to 197.2X (mean = 33.4X; median = 10.5X; **Supplementary Table S1**). For each sample, a concatenation of 5,648 bp from six mitochondrial genes with high sequencing depths (*atp1*, *atp4*, *atp6*, *cob*, *cox3*, and *nad9*) was selected for phylogenetic analyses. For the mtDNA data, there were 130 variable sites not including the outgroup samples. For the nrDNA, we assembled the complete cistron sequence for all samples, including a portion of the external transcribed spacer (*ETS*) and non-transcribed spacer (*NTS*) regions. The length of the alignment was 6,415 bp, with sequencing depths between 885.8X and 7,584.7X and a mean depth of 2,765.2X (median = 2,046.7X; **Supplementary Table S1**). There were 257 variable sites not including the outgroup samples in the nrDNA alignment.

SNP Validation

In order to confirm some of the next-generation sequencing SNPs detected in the alignments, we amplified and Sanger sequenced 30 individuals for the *trnK* gene of the cpDNA genome (alignment length = 806 bp). We Sanger sequenced 16,648 bp to compare with the next-generation sequences. A total of 16,502 bp (99.1%) agreed with those obtained during the next-generation sequencing and assembly process. Ninety-four base pairs (0.56%) correspond to ambiguities in the next-generation sequencing that were only partially confirmed because we could not amplify all individuals pooled for that population sample (i.e. we detected one of the bases that cause the ambiguity, but not the other). The remaining 52 bp (0.31%) could not be confirmed due to recalcitrant amplification of some DNA samples. Of the 60 putative SNPs present in the *trnK* gene, 3 SNPs (5.0%) were completely validated, 47 SNPs (78.3%) had ambiguities that were partially confirmed, and 10 SNPs (16.7%) could not be verified because of failed PCR reactions. Sequences obtained from Sanger sequencing did not reveal any incongruences with those obtained during the next-generation sequencing.

We also amplified and Sanger sequenced 28 individuals for the *atp1* gene of the mtDNA genome (alignment length = 753). In total, 15,813 bp of this region were compared with the next-generation sequences. We confirmed the veracity of 15,509 bp (98.1%), whereas 213 bp (1.35%) were partially validated (i.e. one of the two possible bases were identified at an ambiguous position), and 91 bp (0.58%) could not be confirmed due to failed PCR reactions. Of the 125 putative SNPs present in this gene, 11 SNPs (8.8%) were confirmed, 98 SNPs (78.4%) contained at least one ambiguity that was partially validated, and 16 SNPs (12.8%) could not be verified because of failed PCR reactions. Sequences obtained from Sanger sequencing did not reveal incongruences with those obtained during the next-generation sequencing.

The variability within populations (i.e. ambiguities resulting of pooling five individuals in each population) in both the *trnK* and the *atp1* regions was assessed by amplifying and Sanger sequencing two or more individuals per population. When two individuals per population were sequenced, genomic polymorphism was verified in 2 of 21 cases (9.5%), and ambiguities found in next-generation sequencing were partially validated in the remaining cases (90.5%). When three or four individuals per population were sequenced, we confirmed the population genomic polymorphism in 24 of 29 cases (82.8%) and performed partial validations of the ambiguities found in next-generation sequencing in the remaining cases (17.2%). Failed PCR reactions precluded our ability to sequence all five individuals of any population.

Phylogenetic Analyses

ML and Bayesian phylogenetic analysis of the cpDNA genome showed mostly congruent topologies (**Figure 3A** and **Supplementary Figure S1A**). In the ML analysis, 17 of 27 (63%)

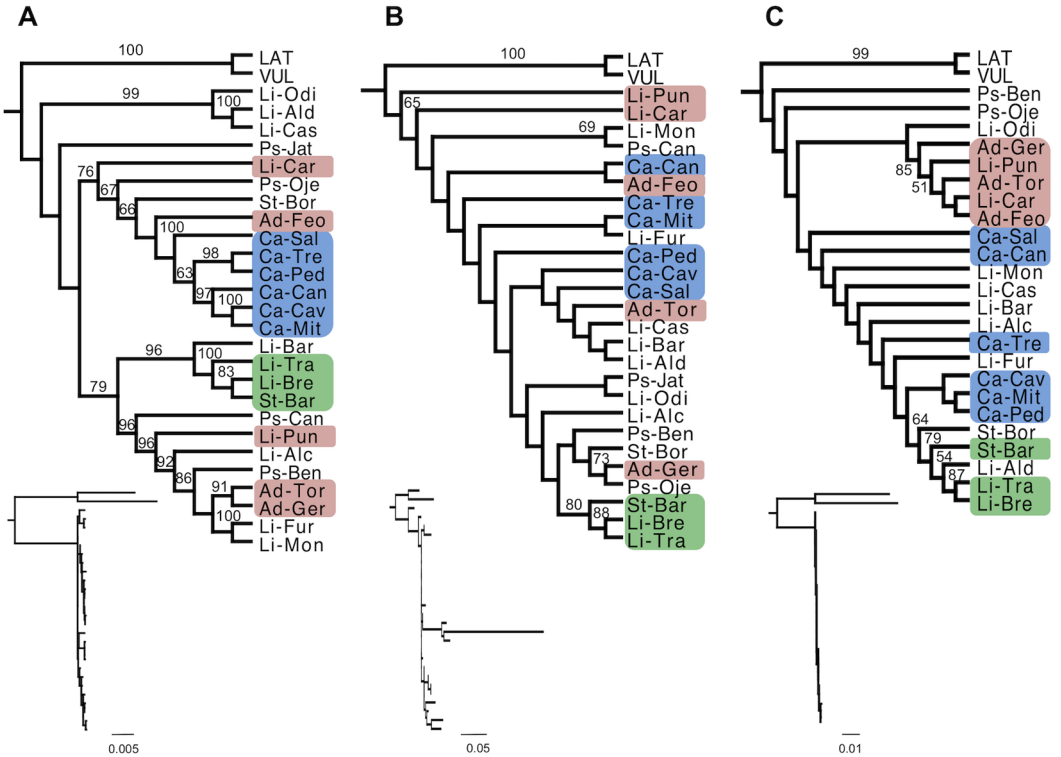

FIGURE 3 | Phylogenetic relationships within the Section *Psammophilae* using maximum likelihood (ML) estimation. Phylogenetic relationships were determined from the complete chloroplast genomes **(A)**, six coding regions of the mitochondrial genome **(B)**, and complete nuclear ribosomal cistron with a partial ETS **(C)**. Sequences were aligned, rooted with two outgroups, and analyzes under maximum likelihood (ML) phylogenetic methods (RAxML). Numbers above branches represent bootstrap support (BS > 50 are displayed). Cladograms show relationships among taxa, while branch lengths are displayed in the inset phylograms. Branches are drawn proportional to the number of substitutions per site (see scale bar). Blue, green, and red squares represent populations from the Balearic Islands and the southern (Cadiz province) and southeast (Almeria province) of the Iberian Peninsula, respectively. Species and population codes are shown in **Table 1**.

internal branches were moderately or highly supported, while in the Bayesian analysis, 20 of 27 (74%) branches showed strong support. There are very few geographic patterns, and most species are not reciprocally monophyletic. An exception to this overall pattern is the six populations of *S. cambessedesii* from the Balearic Islands which form a strongly supported clade (BS = 100; **Figure 3A**). On the mainland, three geographically adjacent populations of *S. littorea* and *S. stockenii* (Li-Bre, Li-Tra, and St-Bar; see species and population codes in **Table 1**) from Cadiz province, in the southern end of the Iberian Peninsula, form another strongly supported clade (BS = 100). The network analysis of the cpDNA data produced a largely unresolved "starburst" with some differentiation of the Balearic Island samples combined with some mainland samples (**Supplementary Figure S2A**).

Topologies for mtDNA-based trees were mostly congruent for ML and Bayesian phylogenetic analysis (**Figure 3B** and **Supplementary Figure S1B**), yet only a few internal branches were confidentially resolved. In the ML analysis, 4 of 27 (15%) internal branches were moderately or highly supported, while in the Bayesian analysis, 6 of 27 (22%) branches showed moderate or strong support. The mtDNA-based trees did not show any clear phylogeographic pattern. Neither island nor mainland populations clustered geographically, except for the three adjacent populations of Cadiz province (Li-Bre, Li-Tra, and St-Bar) that formed a well-supported clade (BS = 88). The network analysis of the mtDNA data produced largely unresolved splits that do not correspond to clear taxon or geographical delimitations. In addition, the *S. littorea* from Barra (Li-Bar) had an exceptionally long terminal branch (**Supplementary Figure S2B**).

For the nrDNA genome, ML and Bayesian phylogenetic analysis were mostly congruent (**Figure 3C** and **Supplementary Figure S1C**), although very few internal branches were resolved. In the ML analysis, 4 of 27 (15%) internal branches were moderately or highly supported, while in the Bayesian analysis, only 4 of 27 (15%) branches showed moderate or strong support. Phylogenetic relationships in the nrDNA-based trees did not reflect either taxonomic boundaries or biogeographic patterns. The same three adjacent populations of Cadiz province (Li-Bre, Li-Tra, and St-Bar), together with a southern population of *S. littorea* (Li-Ald) separated approximately 65 km from these populations, formed a moderate supported clade (BS = 79). A strongly supported clade (BS = 85) emerged in the southeast of the Iberian Peninsula (Almeria province), composed by five adjacent populations: three populations of *S. adscendens* (Ad-Feo, Ad-Ger, and Ad-Tor) and two populations of *S. littorea* (Li-Car and Li-Pun). The populations of *S. psammitis* from

TABLE 1 | Sample localities (populations of each species are ordered by name) and distance to the closest species.

Code	Species	Locality information	Latitude (°)	Longitude (°)	Distance to closest spp. (km)	Closest spp.
Ad-Feo	*Silene adscendens*	Spain, Almería, Los Feos	37.013444	–2.029278	13	*S. littorea*
Ad-Ger	*S. adscendens*	Spain, Almería, Gerjal	37.083361	–2.507861	20	*S. psammitis*
Ad-Tor	*S. adscendens*	Spain, Almería, Los Toros	36.822639	–2.043222	19	*S. littorea*
Ca-Can	*Silene cambessedesii*	Spain, Formentera, Canyes	38.729528	1.451861	184	*S. littorea*
Ca-Mit	*S. cambessedesii*	Spain, Formentera, Mitjorn	38.684389	1.467500	184	*S. littorea*
Ca-Sal	*S. cambessedesii*	Spain, Formentera, Ses Salines	38.746806	1.432889	183	*S. littorea*
Ca-Cav	*S. cambessedesii*	Spain, Ibiza, Cavallet	38.848139	1.401056	184	*S. littorea*
Ca-Ped	*S. cambessedesii*	Spain, Ibiza, Sa Pedrera	38.970028	1.261111	180	*S. littorea*
Ca-Tre	*S. cambessedesii*	Spain, Ibiza, Punta des Trencs	38.969194	1.270722	179	*S. littorea*
Li-Mon	*Silene littorea*	Portugal, Faro, Monte Clérigo	37.341174	–8.852668	95	*S. psammitis*
Li-Cas	*S. littorea*	Portugal, Lisboa, Cascais	38.702153	–9.473942	85	*S. psammitis*
Li-Alc	*S. littorea*	Portugal, Setúbal, Alcácer do Sal	38.485790	–8.903009	25	*S. psammitis*
Li-Fur	*S. littorea*	Spain, A Coruña, Furnas	42.638420	–9.039037	210	*S. psammitis*
Li-Car	*S. littorea*	Spain, Almería, Carboneras	36.962500	–1.899722	13	*S. adscendens*
Li-Pun	*S. littorea*	Spain, Almería, Punta Entinas	36.710261	–2.639618	7	*S. adscendens*
Li-Bre	*S. littorea*	Spain, Cádiz, Breña	36.189620	–5.949146	7	*S. stockenii*
Li-Tra	*S. littorea*	Spain, Cádiz, Trafalgar	36.182506	–6.034710	13	*S. stockenii*
Li-Odi	*S. littorea*	Spain, Huelva, Odiel	37.164706	–6.919111	45	*S. psammitis*
Li-Ald	*S. littorea*	Spain, Málaga, Aldea Beach	36.332278	–5.239083	16	*S. psammitis*
Li-Bar	*S. littorea*	Spain, Pontevedra, Barra	42.259707	–8.840256	180	*S. psammitis*
Ps-Can	*Silene psammitis*	Spain, Ávila, Candeleda	40.215608	–5.247733	300	*S. littorea*
Ps-Jat	*S. psammitis*	Spain, Granada, Játar	36.916194	–3.905028	40	*S. littorea*
Ps-Ben	*S. psammitis*	Spain, Málaga, Benahavis	36.511000	–5.035750	27	*S. littorea*
Ps-Oje	*S. psammitis*	Spain, Málaga, Ojen	36.592972	–4.857389	18	*S. littorea*
St-Bar	*Silene stockenii*	Spain, Cádiz, Barca de Vejer	36.247929	–5.914718	7	*S. littorea*
St-Bor	*S. stockenii*	Spain, Cádiz, Bornos	36.818347	–5.767805	65	*S. psammitis*

Candeleda and Játar (Ps-Can, Ps-Jat) are not present in the nrDNA phylogenetic analyses because we obtained paralogous sequences during the assemblies (the average distance to the mean patristic distance within the ingroup was 87.5%). We tried to recover the orthologous copies by remapping the reads using the consensus sequence of *S. psammitis* from Benahavis (Ps-Ben) as a reference, but this failed to produce homologous sequences. The network analysis of the nrDNA data produced a set of relationships largely congruent with the phylogenetic results described above. There were four clusters of samples that were largely biogeographically aligned—"Northwest–West Iberian Peninsula," "South–Southwestern Iberian Peninsula," "Southeast Iberian Peninsula," and "Balearic Islands" (**Supplementary Figure S2C**).

Gene Tree–Species Tree Reconciliation

For this and all subsequent analyses, we focus on the cpDNA genome because this organelle traces colonization events (although chloroplast capture could entangle phylogenetic relationships), which are particularly informative to clarify evolutionary relationships at the intra- and interspecific level, whereas nrDNA reflects both seed and pollen gene flows. The species tree obtained from *BEAST analysis of the cpDNA genome clustered together the five species of the *Silene* section *Psammophilae* (*PP* = 1; **Figure 4A**). The island populations of *S. cambessedesii* are strongly supported (*PP* = 0.99) as sister clade to the remaining mainland species, which showed weakly supported relationships among them. The species trees from DensiTree highlight the network-like relationships within the section, especially among mainland taxa (**Figure 4B**). The most recent common ancestor of the section emerged approximately 2.34 Ma (95% HPD: 0.05–7.48 Ma), while the origins of *S. adscendens*, *S. littorea*, *S. psammitis*, and *S. stockenii* are more recent and very similar to one another [~1.33 (95% HPD: 0.02–4.13) Ma] (**Figure 4A**).

Population Genetic Structure

We used DAPC of the plastid genome to investigate species affinities and to look for traces of organellar introgression. DAPC analysis revealed that the probability of membership to the assigned species priors was unequivocal for *S. cambessedesii* (100%), but variable for mainland populations, ranging from 12.7% to 100% (**Figure 2**). *S. cambessedesii* samples from the Balearic Islands were mostly separated from mainland populations along the first two retained principle component axes of the DAPC (representing 66.5% of the variation), whereas mainland populations showed largely overlapping 95% inertial ellipses (**Figure 5A**).

When using geography as priors, the probability of correct assignment to samples with regard to geographic location was again unequivocal for Balearic populations of *S. cambessedesii* (100%) and varied widely for the mainland populations (16.2–100%). *S. cambessedesii* and *S. psammitis* population from Candeleda (named as east and interior mainland locations, respectively, in **Figure 5B**) are clearly separated from the populations from the northwest, west, south, and southeast of the Iberian Peninsula, which overlapped along the first two retained principle component axes of the DAPC (**Figure 5B**).

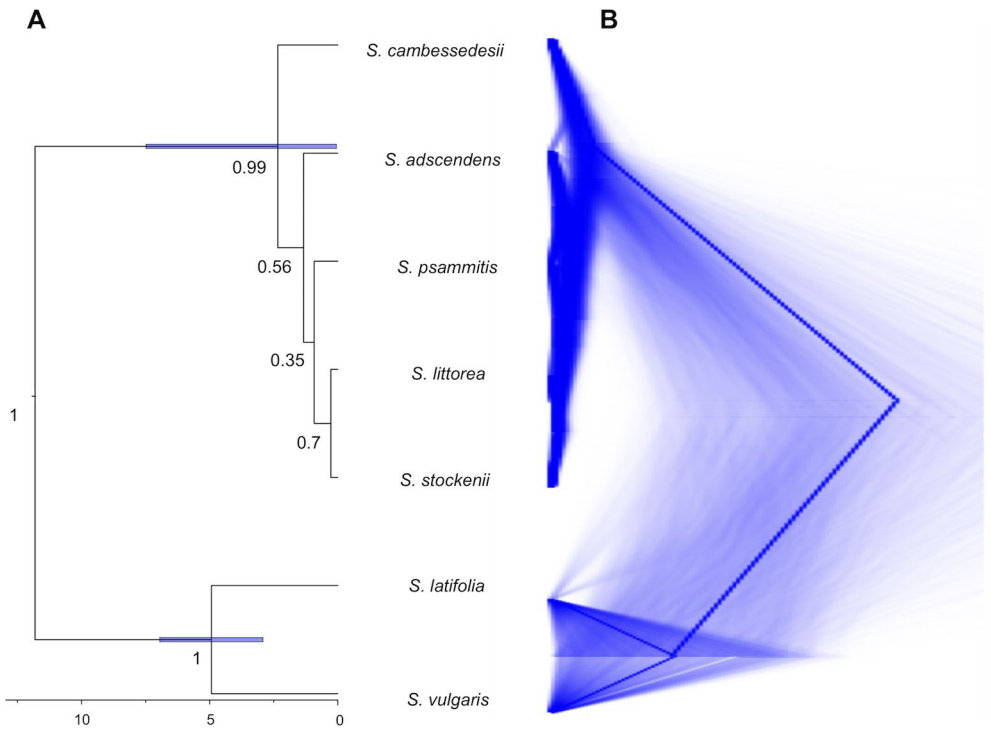

FIGURE 4 | Species tree reconciliation analysis. **(A)** *BEAST multispecies coalescent with posterior probabilities over branches and node age estimates (95% HPD) with posterior probability limit set up to 0.90 (strong support). **(B)** DensiTree visualization of the posterior distribution of gene trees obtained from *BEAST.

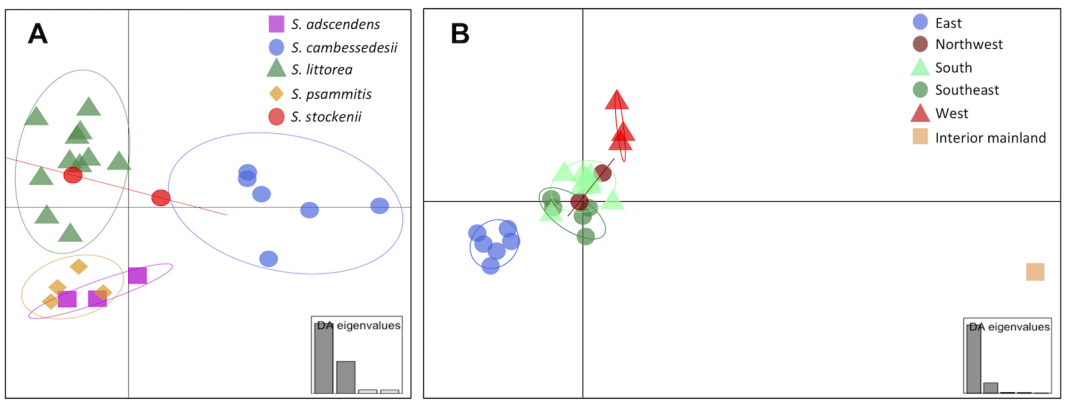

FIGURE 5 | Scatter plot of Discriminant Analysis of Principal Components (DAPC) when using species and location categories as priors. In the scatter plot using species categories as priors **(A)**, each species is represented by different symbols and colors (*S. adscendens*, pink square; *S. cambessedesii*, blue circle; *S. littorea*, green triangle; *S. psammitis*, orange diamond; *S. stockenii*, red circle). In the scatter plot using geographic location categories as priors **(B)**, each location is represented by different symbols and colors (northwest, dark red circles; west, red triangles; south, green triangles; southeast, dark green circles; east, blue circles; interior mainland, brown square). The 95% inertial ellipses around each cluster represent the variance of the two first principal components of the DAPC. The insets represent the relative magnitude of the eigenvalues of the first four and five principal components, respectively.

The BAP analysis, performed to determine if the cpDNA data could be subdivided, conclusively found that there is just one genetic cluster. The log likelihood values for K_1 were much higher than those for other partitions ($K_1 = -37{,}598.87$, $K_2 = -39{,}489.59$, $K_3 = -39{,}576.70$, $K_4 = -40{,}069.07$).

Partial Mantel tests were performed to test for isolation-by-distance and showed that pairwise genetic distances were very low for island populations, but variable for mainland populations. Partial Mantel test showed spatial correlation of patristic distances when all populations were considered ($P < 0.001$, $r = 0.35$). However, when the partial Mantel tests were conducted separately for mainland and island populations, there was no significant correlation among mainland population ($P = 0.27$, $r = 0.07$), but there was marginally significant

correlation among island populations (P = 0.079, r = 0.49) (**Supplementary Figure S3**).

We conducted AMOVA using the cpDNA data set in order to determine the distribution of genetic variation within and among species. There was a moderate level of genetic differentiation among species (F_{ST} = 0.23; P < 0.001); most of the genetic variation was concentrated within species (77.3%) compared to among species (22.7%; **Table 2**). However, when the AMOVA was restricted to mainland species, 100% of the genetic variation lies among populations within species (**Table 3**). Finally, when the AMOVA was restricted to island populations, we did not find genetic differentiation between the two Balearic Islands (**Table 4**).

Similarity of *S. cambessedesii* Plants From Mainland and Balearic Populations

Sanger sequencing of four loci (*trnK* and *ycf1* from the cpDNA, *atp1* from the mtDNA, and ITS from nrDNA) from 20 seeds from a mainland population of *S. cambessedesii* (Almenara) generated 2,993 bp, which were compared to the Balearic Island populations sequenced using next-generation. Nearly all the sequences from the Almenara population were identical to those from the Balearic populations (99.2%). For the 23 SNPs (0.8%) that contained ambiguities in one or more of the Balearic populations, the Almenara individuals had one of the bases that causes the ambiguity. Two SNPs were exclusively found in the ITS1 and ITS2 regions of sequences from Balearic populations. SNPs detected in both island and mainland populations of *S. cambessedesii* were also present in several populations from the southeast and the southern end of the Iberian Peninsula. One SNP in the *ycf1* region was exclusively found in all *S. cambessedesii* samples.

In a ML analysis of the *trnK* fragment (cpDNA) including the mainland *S. cambessedesii* samples, only 3 of 28 (10.7%) internal branches were strongly supported (BS = 100) (**Supplementary Figure S4A**). The Almenara population showed a weak relationship with two populations of *S. cambessedesii* from the Balearic Islands, but also with three mainland populations from the south and the southeastern parts of the Iberian Peninsula. In the *ycf1*-based tree (cpDNA), 3 of 28 (10.7%) internal branches showed strong support (**Supplementary Figure S4B**). The Almenara population formed a cluster with all populations of *S. cambessedesii* from the Balearic Islands but was only moderately supported (BS = 62). Within this cluster, only the relationship between Almenara and one population from Ibiza (Ca-Cav) was strongly supported (BS = 97).

In the topology of the *atp1*-based tree (mtDNA), only 2 of 28 (7.1%) internal branches had a moderate or strong support (**Supplementary Figure S4C**). *S. cambessedesii* from Almenara and the Balearic Islands clustered together with three more mainland populations of other species, showing moderate support (BS = 75).

The ML analysis of the ITS region (nrDNA) revealed that 5 of 28 internal branches (17.9%) have moderate to strong support (**Supplementary Figure S4D**). All populations of *S. cambessedesii* from the Balearic Islands formed a very weakly supported clade (BS = 61), but plants from Almenara seemed to be more closely related with other mainland populations.

DISCUSSION

This study sought to reconstruct the relationships among the five species in *Silene* section *Psammophilae*, yet neither the

TABLE 2 | Analysis of molecular variance (AMOVA) testing genetic subdivision among all species.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	Statistics	*P*
Among species	4	1,546.5	47.7	22.7	F_{ST} = 0.23	<0.001
Within species	21	3,402.5	162.0	77.3		
Total	25	4,949.0	209.7			

TABLE 3 | AMOVA testing genetic subdivision among mainland species.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	Statistics	*P*
Among species	3	552.9	–3.61	0	F_{ST} = –0.018	0.58
Within species	16	3,189.4	199.3	100		
Total	19	3,742.3	195.7			

TABLE 4 | AMOVA testing genetic subdivision between islands.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	Statistics	*P*
Among islands	1	42.2	–1.89	0	F_{ST} = –0.04	0.71
Within islands	4	191.3	47.8	100		
Total	5	233.5	45.9			

complete chloroplast genome, nor the complete nrDNA cistron, nor a portion of the mitochondrial genome supported reciprocal monophyly of any of the species except the Balearic Island populations of *S. cambessedesii*. In the following sections, we discuss the potential causes for the incongruence among these three loci and between the DNA sequences and morphology-based species boundaries. Finally, we discuss the utility of genome skimming to obtain next-generation sequence data from three genomes for phylogeographic inference.

Phylogenetic Relationships and Hybridization in Iberian *Silene*

DNA sequence data from all three genomes support a single common ancestor of *Silene* section *Psammophilae*, yet these data did not resolve the phylogenetic relationships therein. Most of the relationships across all three trees are poorly resolved, and a few results even indicate strongly supported, yet incongruent, relationships among the three genomes. Often, geography was a better predictor of relatedness than either morphology-based species boundaries or edaphic preferences of the species. The cpDNA analysis revealed a monophyletic clade formed by the six Balearic populations and another clade of three Cadiz populations representing two distinct species, whereas nrDNA analysis showed a clade formed by five Almeria populations representing two distinct species. Clearly, these three genomes reveal a complex evolutionary history of these species.

Incongruence among gene trees is frequently attributed to hybridization and/or incomplete lineage sorting (Frajman et al., 2009a). The lack of genetic differentiation found in the AMOVA analysis indicates that mainland species have not diverged genetically, probably because of a history of gene flow among them. Hybridization and introgression are common in *Silene* (Frajman and Oxelman, 2007; Frajman et al., 2009a; Rautenberg et al., 2010; Petri et al., 2013). Interspecific hybridization is often more common between closely related species (Mallet, 2005; Widmer et al., 2009), but in the genus *Silene*, hybridization has also been reported among distantly related species (Petri et al., 2013). The phylogenetic proximity and recent time of divergence (around 2.34 Ma) of species of section *Psammophilae* fall well within the possibility of interspecific crossability in this promiscuous genus, yet no one has reported the existence of interspecific hybrids. However, populations at the overlapping margins of the geographic distributions of distinct species have been noted to have intermediate morphological traits (EN personal observations).

In spite of the widespread incongruence between morphology-based species boundaries and phylogenetic relationships, we found occasional evidence of geographic patterning within Section *Psammophilae*. In addition to the monophyly of Balearic samples, in the southern end of the Iberian Peninsula (Cadiz), the proximity of one population of *S. stockenii* (St-Bar) to two populations of *S. littorea* (Li-Bre and Li-Tra), separated by only a few kilometers, may reflect an increased likelihood of gene flow. This could explain the largely overlapping ellipses defining these two species in the DAPC analysis. Similarly, in the nrDNA phylogenetic analysis, five geographically adjacent populations of *S. adscendens* and *S. littorea* from the southeastern part of the Iberian Peninsula cluster together (Ad-Feo, Ad-Ger, Ad-Tor, Li-Car, and Li-Pun). DAPC results show mixed genetic heritage in two of these populations (Ad-Feo and Li-Car) that contrasts with the other three genetically unique populations (Ad-Ger, Ad-Tor, and Li-Pun), which seem to become more isolated in the eastern part of the Baetic System. Our results suggest that, in some cases, geographic proximity of morphologically distinct species increases the probability of interspecific hybridization or introgression. Several studies have revealed that genetic diversity in *Silene* is more reflective of the geography than the taxonomy (e.g. Frajman et al., 2009a; Durović et al., 2017). In addition, genetic admixture between species highly differentiated for numerous phenotypic and ecological traits has been reported in some *Silene* species (e.g. *S. latifolia* and *S. dioica;* Minder et al., 2007; Hathaway et al., 2009), but also in many other genera (e.g. *Populus alba* and *Populus tremula*; Lexer et al., 2010). In the same way, even if the species of the Section *Psammophilae* display morphological variation among them and distinct edaphic affinities, these differences do not seem to preclude gene flow.

Alternatively, the lack of correlation between taxonomy and genetic structure in species of the Section *Psammophilae* could indicate that *Silene* section *Psammophilae* consists of a single species, or maybe two species if we consider that *S. cambessedesii* is genetically distinct from the mainland species. This would appear to be the most likely conclusion from the *BEAST analysis, which generates a largely unresolved tree (**Figure 4B**). Under this scenario, the existing geographically structured variation would have to be a product of genetic drift, phenotypic plasticity, and/or local adaptation. Due to the consistency of the morphological and ecological differences among these lineages, including important phenotypic traits for species identification in Caryophyllaceae such as floral morphology and seed-coat ornamentation (Chater and Walters, 1964; Talavera, 1990), we prefer to treat them as distinct species which we now know harbor cpDNA and portions of the mtDNA and nrDNA with evolutionary histories that are largely incongruent with the morphological boundaries. Additional markers spanning the nuclear genome (e.g. RAD-Seq, comparative transcriptomics, or whole genome sequencing) could be used to test for phylogenetic signal amidst a history of introgressive hybridization among species of the Section *Psammophilae*.

Geographic Isolation Promotes Genetic Differentiation of Balearic Populations

The colonization of islands by one or a few individuals can lead to the fixation of genetic variation in contrast to large, contiguous continental populations (Frankham, 1997; Franks, 2010). The Balearic populations of *S. cambessedesii* may have fixed genetic differences due to founder effects following the colonization of these Mediterranean islands by a small number of individuals and/or subsequent population bottlenecks (Barton and Charlesworth, 1984; Carson and Templeton, 1984; Ellstrand and Elam, 1993; Orsini et al., 2013). The AMOVA analyses confirmed moderate genetic differentiation of this species with respect to mainland species of Section *Psammophilae*, but also showed the

genetic uniformity of all Balearic populations. Genetic diversity in island populations may be influenced by numerous factors, but physical characteristics, such as the distance to other islands and the mainland, are probably one of the most important (García-Verdugo and Fay, 2014; Stuessy et al., 2014). For instance, proximity of California Channel Islands to each other and the mainland precludes island isolation as measured by the genetic diversity of endemic *Acmispon* (McGlaughlin et al., 2014). In contrast, the much larger geographical isolation of Hawaiian silverswords is an impediment to gene flow among islands and to any distant continents (McGlaughlin and Friar, 2011). Ibiza and Formentera are close enough to each other for dispersal to allow repeated gene flow between islands. Moreover, during the Pleistocene glaciations, Ibiza and Formentera formed a single large island as a consequence of sea level drop that allowed the contact among previously isolated populations of *S. cambessedesii* (Rodríguez et al., 2013; Chueca et al., 2015). Therefore, the weak genetic differentiation between islands we observe today is likely due to gene flow facilitated by historical land bridges or island proximity. By contrast, the islands remained isolated from the mainland during this time.

The presence of many endemic species in the Balearic Islands has been frequently explained by colonization events across land corridors that connected the archipelago to the Iberian Peninsula during the Messinian Salinity Crisis (~5.5 Ma) (e.g. Garnatje et al., 2013; Chueca et al., 2015). Since the origin of Section *Psammophilae* seems to be more recent, the presence of *S. cambessedesii* in the Balearic Islands could not be explained by stepwise dispersal across these land corridors. Although *S. cambessedesii* lacks any obvious dispersal mechanisms, other species in this genus are capable of long-distance dispersal (Giles and Goudet, 1997; Eggens et al., 2007; Gussarova et al., 2015). Seed dispersal of coastal species without obvious dispersal mechanisms has been commonly explained by the accidental ingestion of seeds by granivorous birds, seed transportation in the plumage of birds and in mud attached to their legs and feet, or by water dispersal (Carlquist, 1967; Fridriksson, 1975; Richardson et al., 2000). We thus suggest that a single or very few long-distance dispersal events allowed *S. cambessedesii* to colonize the coasts of the Balearic Islands where they became isolated from the rampant interspecific hybridization on the Iberian Peninsula.

The existence of additional populations of *S. cambessedesii* on the mainland might be explained by dispersal from the islands back to the mainland following the genetic differentiation of this species on the Balearic Islands. The genetic diversity in many species distributed on both mainland and islands is highly influenced by gene flow between populations on both sides of water barriers (García-Verdugo et al., 2010). For instance, gene flow between populations of the Japanese shrub, *Weigela coraeensis,* from the Izu Peninsula and the adjacent northern Izu Islands explains why there is genetic differentiation with respect to more isolated populations from the southern islands (Yamada and Maki, 2012). Similarly, several studies have described an east–west geographical pattern in the Canary Islands in which populations from the eastern islands were genetically more similar to mainland taxa (e.g. García-Verdugo et al., 2009). *Silene hifacensis,* restricted to the east coast of the Iberian Peninsula and Ibiza, lacks any clear genetic differentiation between populations at both sides of the water barrier (Prentice et al., 2003). Gene flow between island and mainland populations of *S. cambessedesii* may explain the presence of a divergent nucleotide in the *ycf1* region restricted to this species. However, the remaining divergent SNPs found in *S. cambessedesii* sequences are also shared with several populations from the southeast and the southern end of the Iberian Peninsula. The larger effective population sizes of these mainland populations may continue to harbor genetic variation from before the dispersal event to the archipelago and/or have acquired variation more common in the mainland populations due to hybridization and introgression. Thus, our results suggest that genetic traces observed in DNA sequences of Almenara individuals are the result of at least one dispersal event from the islands back to the mainland, followed by introgression with mainland populations. Nevertheless, future detailed phylogeographic studies applying additional nuclear markers across as many mainland and island populations as possible will be necessary to further investigate the colonization history of *S. cambessedesii* in the Balearic Islands.

On the Relative Accuracy of Reference-Guided Assembly of Each Genome

Reference-guided genome skimming for data from three genomes was sufficient to obtain the complete nuclear ribosomal unit and nearly complete plastome in all samples. The abundance of nrDNA and the high proportion of plastids per nuclear genome make them especially amenable for genome skimming and reference-guided assembly, even when there is no closely related genome to use as a reference (Straub et al., 2011; Straub et al., 2012). On the other hand, mitochondrial assemblies were largely incomplete and mainly restricted to coding regions, similar to other assemblies where mtDNAs were recovered (Malé et al., 2014; Ripma et al., 2014). Despite the abundance of this organelle in genomic DNA, obtaining mitochondrial genomes was challenging due to their complexity, variability, and frequent structural rearrangements in plants (Palmer and Herbon, 1988; Sugiyama et al., 2005; Knoop et al., 2011).

In some species of *Silene*, the mitochondrial genomes have a complex multichromosomal structure with large variations in genome sizes (Sloan et al., 2009), including the gain or loss of entire chromosomes in some species (Wu et al., 2015), and extremely variable substitution rates within and among species (Sloan et al., 2008; Sloan et al., 2009). In the same way, mitochondrial substitution rates also could vary among species of Section *Psammophilae*. In fact, the branch length of the *S. littorea* from Barra (Li-Bar) in the mtDNA-based tree is approximately 10 times longer than its sister taxon (*S. psammitis*) and other populations of the same species. If these differences in branch length are explained by an exceptional variation in mitochondrial substitution rates, and not because of an inaccurate assembly, caution would be needed when using mitochondrial sequence for reconstructing phylogenetic relationships (Felsenstein, 1978). Larger fragments of the mitochondrial genome, including introns and intergenic regions, will be necessary to further examine phylogenetic and biogeographic analyses using this genome (Straub et al., 2012; Bock et al., 2014). Thus, genome skimming is

an efficient approach to generate the majority of the chloroplast genome, nrDNA cistron, and some mitochondrial coding sequences, yet even this cannot overcome the complex, recent, reticulate evolutionary history of *Silene* section *Psammophilae*.

CONCLUSIONS

In this study, we skimmed the chloroplast genome, complete nrDNA, and portions of the mitochondrial genome, yet this was largely insufficient to reconstruct the complex evolutionary history of the members of *Silene* section *Psammophilae*. Except in the presence of substantial biogeographic barriers (e.g. the Balearic Islands), the highly reticulated evolutionary histories of young lineages and pervasive hybridization will remain challenging, even with massive amounts of sequence data at hand.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the Genbank (MN365968 - MN365993, MN334914 - MN334939, MN334784 - MN334809, MN334810 - MN334835, MN334836 - MN334861, MN334862 - MN334887, MN334888 - MN334913, MN325944 - MN325960).

AUTHOR CONTRIBUTIONS

EN, MB, and JW conceived of the experiments. MB, IC-S, and EN carried out the sampling. IC-S and JW performed the DNA extraction. JW and JV performed the assembly and the alignment of sequences. JW and JV performed the Sanger sequencing for SNP validation. Phylogenetic analyses were conducted by JV and JW. JV and JW wrote the article, with assistance from all coauthors. All authors read and approved the final manuscript.

FUNDING

This work was supported with FEDER funds by the Spanish Government MINECO projects (CGL2009-08257, CGL2012-37646, and CGL2015-63827-P) and the Predoctoral Training Program grants to IS (BES-2010-031073) and JV (BES-2013–062610).

ACKNOWLEDGMENTS

The authors thank Charles Nicolet and the University of Southern California's Epigenome Center for library preparation and conducting the sequencing. The Department of Biology of Santa Clara University supported IC-S and JV during two research stays to conduct an important portion of these analyses. D. Baker provided computational support at Santa Clara University. The Servicio de Vida Silvestre-CIEF in Valencia (Spain) provided us the material from the *S. cambessedesii* population of Almenara. We also thank to M. Escudero, E. Maguilla, S. Martín-Bravo, and V. Valcárcel for their contribution with data analysis and helpful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01466/full#supplementary-material

REFERENCES

Arnold, M. L. (1997).*Natural hybridization and evolution*. Oxford, U.K: Oxford University Press.

Baker, H. G. (1948). Stages in invasion and replacement demonstrated by species of *Melandrium*. *J. Ecol.* 36, 96–119. doi: 10.2307/2256649

Balounova, V., Gogela, R., Cegan, R., Cangren, P., Zluvova, J., and Safar, J. (2019). Evolution of sex determination and heterogamety changes in section *Otites* of the genus *Silene*. *Sci. Rep.* 9, 1045. doi: 10.1038/s41598-018-37412-x

Barr, C. M., Keller, S. R., Ingvarsson, P. K., Sloan, D. B., and Taylor, D. R. (2007). Variation in mutation rate and polymorphism among mitochondrial genes of *Silene vulgaris*. *Mol. Biol. Evol.* 24, 1783–1791. doi: 10.1093/molbev/msm106

Barton, N. H., and Charlesworth, B. (1984). Genetic revolutions, founder effects, and speciation. *Annu. Rev. Ecol. Syst.* 15, 133–164. doi: 10.1146/annurev.es.15.110184.001025

Bernasconi, G., Antonovics, J., Biere, A., Charlesworth, D., Delph, L. F., and Filatov, D. (2009). *Silene* as a model system in ecology and evolution. *Heredity (Edinb).* 103, 5–14. doi: 10.1038/hdy.2009.34

Bittkau, C., and Comes, H. P. (2009). Molecular inference of a Late Pleistocene diversification shift in *Nigella* s. lat. (Ranunculaceae) resulting from increased speciation in the Aegean archipelago. *J. Biogeogr.* 36, 1346–1360. doi: 10.1111/j.1365-2699.2008.02003.x

Bock, D. G., Kane, N. C., Ebert, D. P., and ,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

Bouckaert, R., and Heled, J. (2014). DensiTree 2: seeing trees through the forest. *bioRxiv* 10, 012401. doi: 10.1101/012401

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

Carlquist, S. (1967). The biota of long-distance dispersal. V. Plant dispersal to Pacific Islands. *Bull. Torrey Bot. Club* 94, 129–162. doi: 10.2307/2484044

Carson, H. L., and Templeton, A. R. (1984). Genetic revolutions in relation to speciation phenomena: the founding of new populations. *Annu. Rev. Ecol. Syst.* 15, 97–131. doi: 10.1146/annurev.es.15.110184.000525

Casimiro-Soriguer, I. (2015). Sistemas sexuales y polimorfismo de color en *Silene*: una aproximación en la sección *Psammophilae*. Ph.D. Dissertation. Seville: Pablo de Olavide University.

Chater, O. A., and Walters, S. M. (1964). *Silene* in *Flora Europaea*, Vol. 1 . Eds. Tutin,T. G., Heywood,V. H., Burges,N. A., Valentine,D. H., and Walters,S. M. (Cambridge, UK: Cambridge University Press), 158–181.

Chueca, L. J., Madeira, M. J., and Gómez-Moliner, B. J. (2015). Biogeography of the land snail genus *Allognathus* (Helicidae): middle Miocene colonization of the Balearic Islands. *J. Biogeogr.* 42, 1845–1857. doi: 10.1111/jbi.12549

Corander, J., Marttinen, P., Sirén, J., and Tang, J. (2008). Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinf.* 9, 539. doi: 10.1186/1471-2105-9-539

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

Dick, C. A., Buenrostro, J., Butler, T., Carlson, M. L., Kliebenstein, D. J., and Whittall, J. B. (2011). Arctic mustard flower color polymorphism controlled by petal-specific downregulation at the threshold of the anthocyanin biosynthetic pathway. *PloS One* 6, e18230. doi: 10.1371/journal.pone.0018230

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29, 1969–1973. doi: 10.1093/molbev/mss075

Du Pasquier, P. E., Jeanmonod, D., and Naciri, Y. (2017). Morphological convergence in the recently diversified *Silene gigantea* complex (Caryophyllaceae) in the Balkan Peninsula and south-western Turkey, with the description of a new subspecies. *Bot. J. Linn. Soc* 183, 474–493. doi: 10.1093/botlinnean/bow016

Duggen, S., Hoernle, K., Bogaard, P., Van Den, R. L., and Morgan, J. P. (2003). Deep roots of the Messinian salinity crisis. *Nat.* 422, 602–606. doi: 10.1038/nature01551.1

Durović, S., Schönswetter, P., Niketić, M., Tomović, G., and Frajman, B. (2017). Disentangling relationships among the members of the Silene saxifraga alliance (Caryophyllaceae): Phylogenetic structure is geographically rather than taxonomically segregated. *Taxon* 66, 343–364. doi: 10.12705/662.4

Eggens, F., Popp, M., Nepokroeff, M., Wagner, W. L., and Oxelman, B. (2007). The origin and number of introductions of the Hawaiian endemic *Silene* species (Caryophyllaceae). *Am. J. Bot.* 94, 210–218. doi: 10.3732/ajb.94.2.210

Ellstrand, N. C., and Elam, D. R. (1993). Population genetic consequences of small population size: implications for plant conservation. *Annu. Rev. Ecol. Syst.* 24, 217–242. doi: 10.1146/annurev.es.24.110193.001245

Erixon, P., and Oxelman, B. (2008). Reticulate or tree-like chloroplast DNA evolution in *Sileneae* (Caryophyllaceae)?. *Mol. Phylogenet. Evol.* 48, 313–325. doi: 10.1016/j.ympev.2008.04.015

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

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genet.* 131, 479–491.

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

Frajman, B., Eggens, F., and Oxelman, B. (2009a). Hybrid origins and homoploid reticulate evolution within *Heliosperma* (*Sileneae*, Caryophyllaceae) - A multigene phylogenetic approach with relative dating. *Syst. Biol.* 58, 328–345. doi: 10.1093/sysbio/syp030

Frajman, B., Heidari, N., and Oxelman, B. (2009b). Phylogenetic relationships of Atocion and Viscaria (*Sileneae*, Caryophyllaceae) inferred from chloroplast, nuclear ribosomal, and low-copy gene DNA sequences. *Taxon* 58, 811–824. doi: 10.1002/tax.583010

Frajman, B., and Oxelman, B. (2007). Reticulate phylogenetics and phytogeographical structure of *Heliosperma* (*Sileneae*, Caryophyllaceae) inferred from chloroplast and nuclear DNA sequences. *Mol. Phylogenet. Evol.* 43, 140–155. doi: 10.1016/j.ympev.2006.11.003

Frankham, R. (1997). Do island populations have less genetic variation than mainland populations?. *Heredity (Edinb).* 78, 311–327. doi: 10.1038/hdy.1997.46

Franks, S. J. (2010). Genetics, evolution, and conservation of Island plants. *J. Plant Biol.* 53, 1–9. doi: 10.1007/s12374-009-9086-y

Fridriksson, S. (1975).*Surtsey, Evolution of Life on a Volcanic Island*. Butterworth London, UK: John Wiley.

García-Verdugo, C., and Fay, M. F. (2014). Ecology and evolution on oceanic islands: broadening the botanical perspective. *Bot. J. Linn. Soc* 174, 271–275. doi: 10.1111/boj.12154

García-Verdugo, C., Fay, M. F., Granado-Yela, C., De Casas, R. R., Balaguer, L., and Besnard, G. (2009). Genetic diversity and differentiation processes in the ploidy series of *Olea europaea* L.: a multiscale approach from subspecies to insular populations. *Mol. Ecol.* 18, 454–467. doi: 10.1111/j.1365-294X.2008.04027.x

García-Verdugo, C., Forrest, A. D., Fay, M. F., and Vargas, P. (2010). The relevance of gene flow in metapopulation dynamics of an oceanic island endemic, *Olea Europaea* subsp. *Guanchica*. *Evol. (N. Y).* 64, 3525–3536. doi: 10.1111/j.1558-5646.2010.01091.x

Garnatje, T., Pérez-Collazos, E., Pellicer, J., and Catalán, P. (2013). Balearic insular isolation and large continental spread framed the phylogeography of the western Mediterranean *Cheirolophus intybaceus s.l.* (Asteraceae). *Plant Biol.* 15, 166–175. doi: 10.1111/j.1438-8677.2012.00632.x

Giles, B. E., and Goudet, J. (1997). Genetic differentiation in *Silene dioica* metapopulations: estimation of spatiotemporal effects in a successional plant species. *Am. Nat.* 149, 507–526. doi: 10.1086/286002

Greuter, W. (1995). *Silene* (Caryophyllaceae) in Greece: a subgeneric and sectional classification. *Taxon* 44, 543–581. doi: 10.2307/1223499

Gussarova, G., Allen, G. A., Mikhaylova, Y., McCormick, L. J., Mirré, V., and Marr, K. L. (2015). Vicariance, long-distance dispersal, and regional extinction-recolonization dynamics explain the disjunct circumpolar distribution of the arctic-alpine plant *Silene acaulis*. *Am. J. Bot.* 102, 1703–1720. doi: 10.3732/ajb.1500072

Hathaway, L., Malm, J. U., and Prentice, H. C. (2009). Geographically congruent large-scale patterns of plastid haplotype variation in the European herbs Silene dioica and S. latifolia (Caryophyllaceae). *Biol. J. Linn. Soc* 161, 153–170. doi: 10.1111/j.1095-8339.2009.01003.x

Heled, J., and Drummond, A. J. (2010). Bayesian Inference of Species Trees from Multilocus Data using *BEAST. *Mol. Biol. Evol.* 27, 570–580. doi: 10.1093/molbev/msp274

Holland, B. R., Benthin, S., Lockhart, P. J., Moulton, V., and Huber, K. T. (2008). Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evol. Biol.* 8, 202. doi: 10.1186/1471-2148-8-202

Hsü, K. J., Ryan, W. B., and Cita, M. B. (1973). Late Miocene desiccation of the Mediterranean. *Nat.* 242, 240–244. doi: 10.1038/242240a0

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference ofphylogenetic trees. *Bioinforma. Appl. Note* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

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

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

Jombart, T. (2008). *adegenet*: a R package for the multivariate analysis of genetic markers. *Bioinf.* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., and Collins, C. A. (2015). *Tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 2.0.0*. Available at: http://adegenet.r-forge.r-project.org/files/tutorial-dapc.pdf.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: anew method for the analysis of genetically structured populations. *BMC Genet.* 11, 94. doi: 10.1186/1471-2156-11-94

Juan, A., Crespo, M. B., Cowan, R. S., Lexer, C., and Fay, M. F. (2004). Patterns of variability and gene flow in *Medicago citrina*, an endangered endemic of islands in the western Mediterranean, as revealed by amplified fragment length polymorphism (AFLP). *Mol. Ecol.* 13, 2679–2690. doi: 10.1111/j.1365-294X.2004.02289.x

Kane, N., Sveinsson, S., Dempewolf, H., Yang, J. Y., Zhang, D., and Engels, J. M. M. (2012). Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *Am. J. Bot.* 99, 320–329. doi: 10.3732/ajb.1100570

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., Sturrock, S., and Meintjes, P. (2012). *The Geneious 6.0.3 read mapper*. Available at: https://assets.geneious.com/documentation/geneious/GeneiousReadMapper.pdf

Knoop, V., Volkmar, U., Hecht, J., and Grewe, F. (2011). "Mitochondrial genome evolution in the plant lineage," in *Plant mitochondria* (New York, NY: Springer), 3–29. doi: 10.1007/978-0-387-89781-3_1

Krijgsman, W., Hilgen, F. J., Raffi, I., Sierro, F. J., and Wilson, D. S. (1999). Chronology, causes and progression of the Messinian salinity crisis. *Nat.* 400, 652–655. doi: 10.1038/23231

Lexer, C., Joseph, J. A., Van Loo, M., Barbará, T., Heinze, B., and Bartha, D. (2010). Genomic admixture analysis in European *Populus* spp. reveals unexpected patterns of reproductive isolation and mating. *Genet.* 186, 699–712. doi: 10.1534/genetics.110.118828

Malé, P. J. G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., and Engel, J. (2014). Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Mol. Ecol. Resour.* 14, 966–975. doi: 10.1111/1755-0998.12246

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20, 229–237. doi: 10.1016/j.tree.2005.02.010

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27, 209–220.

McGlaughlin, M. E., and Friar, E. A. (2011). Evolutionary diversification and geographical isolation in *Dubautia laxa* (Asteraceae), a widespread member of

the Hawaiian silversword alliance. *Ann. Bot.* 107, 357–370. doi: 10.1093/aob/mcq252

McGlaughlin, M. E., Wallace, L. E., Wheeler, G. L., Bresowar, G., Riley, L., and Britten, N. R. (2014). Do the island biogeography predictions of MacArthur and Wilson hold when examining genetic diversity on the near mainland California Channel Islands ? Examples from endemic *Acmispon* (Fabaceae). *Bot. J. Linn. Soc* 174, 289–304. doi: 10.1111/boj.12122

Médail, F., and Quézel, P. (1997). Hot-spots analysis for conservation of plant biodiversity in the Mediterranean Basin. *Ann. Missouri Bot. Gard.* 84, 112–127. doi: 10.2307/2399957

Médail, F., and Quézel, P. (1999). Biodiversity hotspots in the Mediterranean Basin: setting global conservation priorities. *Conserv. Biol.* 13, 1510–1513. doi: 10.1046/j.1523-1739.1999.98467.x

Minder, A. M., Rothenbuehler, C., and Widmer, A. (2007). Genetic structure of hybrid zones between *Silene latifolia* and *Silene dioica* (Caryophyllaceae): evidence for introgressive hybridization. *Mol. Ecol.* 16, 2504–2516. doi: 10.1111/j.1365-294X.2007.03292.x

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nat.* 403, 853–858. doi: 10.1038/35002501

Naciri, Y., Du Pasquier, P. E., Lundberg, M., Jeanmonod, D., and Oxelman, B. (2017). A phylogenetic circumscription of *Silene* sect. *Siphonomorpha* (Caryophyllaceae) in the Mediterranean Basin. *Taxon* 66, 91–108. doi: 10.12705/661.5

Navarro, A., Ferrer-Gallego, P. P., Ferrando, I., Albert, F. J., Martínez, V., and Escribá, M. C. (2015). Experiencias de conservación activa e *in situ* con *Silene cambessedesii*, especie en peligro de extinción en la Comunidad Valenciana. *Conserv. Veg.* 19, 11–13.

Nieto Feliner, G. (2014). Patterns and processes in plant phylogeography in the Mediterranean Basin. A review. *Perspect. Plant Ecol. Evol. Syst.* 16, 265–278. doi: 10.1016/j.ppees.2014.07.002

Oksanen, J., Blanchet, F., Friendly, M., Kindt, R., Legendre, P., and McGlinn, D. (2018). *vegan: community ecology package. R package version 2.5-2*. https://CRAN.R-project.org/package=vegan.

Okuyama, Y., Fujii, N., Wakabayashi, M., Kawakita, A., Ito, M., and Watanabe, M. (2005). Nonuniform concerted evolution and chloroplast capture: heterogeneity of observed introgression patterns in three molecular data partition phylogenies of Asian *Mitella* (Saxifragaceae). *Mol. Biol. Evol.* 22, 285–296. doi: 10.1093/molbev/msi016

Orsini, L., Vanoverbeke, J., Swillen, I., Mergeay, J., and De Meester, L. (2013). Drivers of population genetic differentiation in the wild: Isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Mol. Ecol.* 22, 5983–5999. doi: 10.1111/mec.12561

Oxelman, B., and Lidén, M. (1995). Generic boundaries in the tribe *Sileneae* (Caryophyllaceae) as inferred from nuclear rDNA sequences. *Taxon* 44, 525–542. doi: 10.2307/1223498

Oxelman, B., Liden, M., and Berglund, D. (1997). Chloroplast *rps*l6 intron phylogeny of the tribe *Sileneae* (Caryophyllaceae). *Plant Syst. Evol.* 206, 393–410. doi: 10.1007/BF00987959

Oxelman, B., Rautenberg, A., Thollesson, M., Larsson, A., Frajman, B., and Eggens, F., et al. (2013). *Sileneae* taxonomy and systematics. Available at: http://www.sileneae.info.

Oxelman, B., Universitet, U., and Rabeler, R. K. (2001). A revised generic classification of the tribe *Sileneae* (Caryophyllaceae). *Nord. J. Bot.* 20, 743–748. doi: 10.1111/j.1756-1051.2000.tb00760.x

Palmer, J. D., and Herbon, L. A. (1988). Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *J. Mol. Evol.* 28, 87–97. doi: 10.1007/BF02143500

Parks, M., Cronn, R., and Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* 7, 84. doi: 10.1186/1741-7007-7-84

Petri, A., and Oxelman, B. (2011). Phylogenetic relationship within *Silene* (Caryophyllaceae) section *Physolychnis*. *Taxon* 60, 953–968. doi: 10.1002/tax.604002

Petri, A., Pfeil, B. E., and Oxelman, B. (2013). Introgressive hybridization between anciently diverged lineages of *Silene* (Caryophyllaceae). *PloS One* 8, e67729. doi: 10.1371/journal.pone.0067729

Popp, M., and Oxelman, B. (2004). Evolution of a RNA polymerase gene family in *Silene* (Caryophyllaceae) - Incomplete concerted evolution and topological congruence among paralogues. *Syst. Biol.* 53, 914–932. doi: 10.1080/10635150490888840

Popp, M., and Oxelman, B. (2007). Origin and evolution of North American polyploid Silene (Caryophyllaceae). *Am. J. Bot.* 94, 330–349. doi: 10.3732/ajb.94.3.330

Prentice, H. C., Malm, J. U., Mateu-Andres, I., and Segarra-Moragues, J. G. (2003). Allozyme and chloroplast DNA variation in island and mainland populations of the rare Spanish endemic, *Silene hifacensis* (Caryophyllaceae). *Conserv. Genet.* 4, 543–555. doi: 10.1023/A:1025603328704

R Core Team. (2016). *R: a language and environment for statistical computing*. Available at: http://www.r-project.org.

Rambaut, A.*FigTree v1.4.2: tree figure drawing tool*, 2014, Available at: http://tree.bio.ed.ac.uk/software/figtree.

Rambaut, A., Suchard, M. A., Xie, D., and Drummond, A. J.*Tracer v1.6*, 2014, Available at: http://beast.bio.ed.ac.uk/Tracer.

Rautenberg, A., Hathaway, L., Oxelman, B., and Prentice, H. C. (2010). Geographic and phylogenetic patterns in *Silene* section *Melandrium* (Caryophyllaceae) as inferred from chloroplast and nuclear DNA sequences. *Mol. Phylogenet. Evol.* 57, 978–991. doi: 10.1016/j.ympev.2010.08.003

Rautenberg, A., Sloan, D. B., Aldén, V., and Oxelman, B. (2012). Phenolic relationships of *Silene multinervia* and *Silene* section *Conoimorpha* (Caryophyllaceae). *Syst. Bot.* 37, 226–237. doi: 10.1600/036364412X616792

Reyment, R. A. (1983). Palaeontological aspects of island biogeography: colonization and evolution of mammals on Mediterranean islands. *Oikos* 41, 299–306. doi: 10.2307/3544089

Richardson, D. M., Allsopp, N., Antonio, C. M. D., Milton, S. J., and Rejma, M. (2000). Plant invasions – the role of mutualisms. *Biol. Rev.* 75, 65–93. doi: 10.1111/j.1469-185X.1999.tb00041.x

Rieseberg, L. H. (1997). Hybrid origins of plant species. *Annu. Rev. Ecol. Syst.* 28, 359–389. doi: 10.1146/annurev.ecolsys.28.1.359

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., and Nakazato, T. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Sci.* 301, 1211–1216. doi: 10.1126/science.1086949

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

Rodríguez, V., Brown, R. P., Terrasa, B., Pérez-Mellado, V., Castro, J. A., and Picornell, A. (2013). Multilocus genetic diversity and historical biogeography of the endemic wall lizard from Ibiza and Formentera, *Podarcis pityusensis* (Squamata: Lacertidae). *Mol. Ecol.* 22, 4829–4841. doi: 10.1111/mec.12443

Ruhsam, M., Rai, H. S., Mathews, S., Ross, T. G., Graham, S. W., and Raubeson, L. A. (2015). Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in *Araucaria*?. *Mol. Ecol. Resour.* 15, 1067–1078. doi: 10.1111/1755-0998.12375

Sloan, D. B., Barr, C. M., Olson, M. S., Keller, S. R., and Taylor, D. R. (2008). Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA. *Mol. Biol. Evol.* 25, 243–246. doi: 10.1093/molbev/msm266

Sloan, D. B., Oxelman, B., Rautenberg, A., and Taylor, D. R. (2009). Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe *Sileneae*. *BMC Evol. Biol.* 9, 260. doi: 10.1186/1471-2148-9-260

Soltis, D. E., and Kuzoff, R. K. (1995). Discordance between nuclear and chloroplast phylogenies in the *Heuchera* group (Saxifragaceae). *Evol. (N. Y).* 49, 727–742. doi: 10.2307/2410326

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

Straub, S. C. K., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., and Weitemier, K. (2011). Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 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

Stuessy, T. F., Takayama, K., López-Sepúlveda, P., and Crawford, D. J. (2014). Interpretation of patterns of genetic variation in endemic plant species of oceanic islands. *Bot. J. Linn. Soc* 174, 276–288. doi: 10.1111/boj.12088

Sugiyama, Y., Watase, Y., Nagase, M., Makita, N., Yagura, S., and Hirai, A. (2005). The complete nucleotide sequence and multipartite organization of

the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. *Mol. Genet. Genomics* 272, 603–615. doi: 10.1007/s00438-004-1075-8

Talavera, S. (1979). Revisión de la sect. *Erectorefractae* Chowdhuri del género *Silene* L. *Lagascalia* 8, 135–164.

Talavera, S. (1990). Silene in *Flora Ibérica*,Vol. 2 . Eds. Castroviejo,S., Aedo, C., Laínz,M., Muñoz Garmendia,F., Nieto Feliner,G., and Paiva, J. (Madrid, Spain: Real Jardín Botánico) 313–406.

Thompson, J. D. (2005).*Plant evolution in the Mediterranean*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780198515340.001.0001

Van der Geer, A., Lyras, G., de Vox, J., and Dermitzakis, M. (2010).*Evolution of island mammals: adaptation and extinction of placental mammals on islands*. Oxford, U.K: Wiley-Blackwell.

Wang, Y., Chen, Q., Chen, T., Tang, H., Liu, L., and Wang, X. (2016). Phylogenetic insights into Chinese *Rubus* (Rosaceae) from multiple chloroplast and nuclear DNAs. *Front. Plant Sci.* 7, 968. doi: 10.3389/fpls.2016.00968

Whittall, J. B., Carlson, M. L., Beardsley, P. M., Meinke, R. J., and Liston, A. (2006). The Mimulus moschatus alliance (Phrymaceae): molecular and morphological phylogenetics and their conservation implications. *Syst. Bot.* 31, 380–397. doi: 10.1600/036364406777585810

Whittall, J. B., Syring, J., Parks, M., Buenrostro, J., Dick, C., and Liston, A. (2010). Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Mol. Ecol.* 19, 100–114. doi: 10.1111/j.1365-294X.2009.04474.x

Widmer, A., Lexer, C., and Cozzolino, S. (2009). Evolution of reproductive isolation in plants. *Heredity (Edinb).* 102, 31–38. doi: 10.1038/hdy.2008.69

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.* 15, 323–354. doi: 10.1111/j.1469-1809.1949.tb02451.x

Wu, Z., Cuthbert, J. M., Taylor, D. R., and Sloan, D. B. (2015). The massive mitochondrial genome of the angiosperm *Silene noctiflora* is evolving by gain or loss of entire chromosomes. *Proc. Natl. Acad. Sci.* 112, 10185–10191. doi: 10.1073/pnas.1421397112

Yamada, T., and Maki, M. (2012). Impact of geographical isolation on genetic differentiation in insular and mainland populations of *Weigela coraeensis* (Caprifoliaceae) on Honshu and the Izu Islands. *J. Biogeogr.* 39, 901–917. doi: 10.1111/j.1365-2699.2011.02634.x

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: 20 December 2019
doi: 10.3389/fpls.2019.01615

Target Capture Sequencing Unravels *Rubus* Evolution

*Katherine A. Carter[1], Aaron Liston[2], Nahla V. Bassil[3], Lawrence A. Alice[4], Jill M. Bushakra[3], Brittany L. Sutherland[5], Todd C. Mockler[6], Douglas W. Bryant[6] and Kim E. Hummer[3]**

[1] Department of Horticulture, Oregon State University, Corvallis, OR, United States, [2] Department of Botany & Plant Pathology, Oregon State University, Corvallis, OR, United States, [3] National Clonal Germplasm Repository, USDA-ARS, Corvallis, OR, United States, [4] Department of Biology, Western Kentucky University, Bowling Green, KY, United States, [5] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, United States, [6] Mockler Lab, Donald Danforth Plant Sciences Center, St. Louis, MO, United States

Rubus (Rosaceae) comprises more than 500 species with additional commercially cultivated raspberries and blackberries. The most recent (> 100 years old) global taxonomic treatment of the genus defined 12 subgenera; two subgenera were subsequently described and some species were rearranged. Intra- and interspecific ploidy levels and hybridization make phylogenetic estimation of *Rubus* challenging. Our objectives were to estimate the phylogeny of 94 taxonomically and geographically diverse species and three cultivars using chloroplast DNA sequences and target capture of approximately 1,000 low copy nuclear genes; estimate divergence times between major *Rubus* clades; and examine the historical biogeography of species diversification. Target capture sequencing identified eight major groups within *Rubus*. Subgenus *Orobatus* and Subg. *Anoplobatus* were monophyletic, while other recognized subgenera were para- or polyphyletic. Multiple hybridization events likely occurred across the phylogeny at subgeneric levels, e.g., Subg. *Rubus* (blackberries) × Subg. *Idaeobatus* (raspberries) and Subg. *Idaeobatus* × Subg. *Cylactis* (Arctic berries) hybrids. The raspberry heritage within known cultivated blackberry hybrids was confirmed. The most recent common ancestor of the genus was most likely distributed in North America. Multiple distribution events occurred during the Miocene (about 20 Ma) from North America into Asia and Europe across the Bering land bridge and southward crossing the Panamanian Isthmus. *Rubus* species diversified greatly in Asia during the Miocene. *Rubus* taxonomy does not reflect phylogenetic relationships and subgeneric revision is warranted. The most recent common ancestor migrated from North America towards Asia, Europe, and Central and South America early in the Miocene then diversified. Ancestors of the genus *Rubus* may have migrated to Oceania by long distance bird dispersal. This phylogeny presents a roadmap for further *Rubus* systematics research. In conclusion, the target capture dataset provides high resolution between species though it also gave evidence of gene tree/species tree and cytonuclear discordance. Discordance may be due to hybridization or incomplete lineage sorting, rather than a lack of phylogenetic signal. This study illustrates

OPEN ACCESS

Edited by:
Juan Viruel, Royal Botanic Gardens, Kew, United Kingdom

Reviewed by:
Wolf L. Eiserhardt, Aarhus University, Denmark
Isabel Larridon, Royal Botanic Gardens, Kew, United Kingdom

***Correspondence:**
Kim E. Hummer
kim.hummer@usda.gov

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

Received: *04 September 2019*
Accepted: *15 November 2019*
Published: *20 December 2019*

Citation:
Carter KA, Liston A, Bassil NV, Alice LA, Bushakra JM, Sutherland BL, Mockler TC, Bryant DW and Hummer KE (2019) Target Capture Sequencing Unravels Rubus Evolution. Front. Plant Sci. 10:1615. doi: 10.3389/fpls.2019.01615

the importance of using multiple phylogenetic methods when examining complex groups and the utility of software programs that estimate signal conflict within datasets.

Keywords: taxonomy, systematics, biogeography, caneberries, genetic resources, plant migration, phylogenomics

INTRODUCTION

The plant genus *Rubus* (Rosaceae), contains a conservative estimate of more than 500 species (Hytönen et al., 2018) and thousands of cultivars. The annual production of the cultivated brambles (raspberries and blackberries), is economically significant for more than 43 countries (FAO, 2019). Crop wild relatives of this genus contribute to broadening the gene pools for breeding programs to improve these nutritious berry crops.

Varying intra- and interspecific ploidy levels (diploid, $2n = 2x = 14$ to dodecaploid, $2n = 12x = 84$, plus aneuploids), and hybridization (Jennings, 1988; Thompson, 1995; Thompson, 1997; Alice et al., 2001; Sochor et al., 2015; Wang et al., 2015) make phylogenetic estimation challenging. Focke's worldwide taxonomic treatment of *Rubus* (Focke, 1910; Focke, 1911; Focke, 1914), defined 12 subgenera (**Table 1**). Subg. *Rubus* (= *Eubatus* Focke), *Idaeobatus*, and *Malachobatus* contain the most species with > 300 species/microspecies for subg. *Rubus*, 88 species for subg. *Idaeobatus* and 92 species for subg. *Malachobatus* in China alone (Jennings, 1988; Lu and Bufford, 2003).

Subg. *Rubus* occurs in the Americas and Europe while *Idaeobatus* is distributed in North America, Europe, Africa and Asia; *Malachobatus* is Asian (Focke, 1910; Focke, 1911; Focke, 1914; Hytönen et al., 2018). Sections *Micranthobatus* and *Lampobatus* were sect. in Focke for species from Australia, Tasmania, and New Zealand (Bean, 1995; Bean, 1997). Some subg. *Dalibarda* species were moved to subg. *Cylactis* (Bailey, 1941). The Flora of China (Lu and Boufford, 2003), which did not consider global taxa, regrouped species into eight sections corresponding to Focke's subgenera of similar names. China is a center of species diversity with 139 endemics (Lu and Bufford, 2003).

Alice and Campbell (1999) published a molecular phylogenetic study that sampled the 12 classic subgenera and species reclassified subsequently in new subgenera described and found that *Anoplobatus*, *Orobatus* and *Rubus*, excluding allopolyploids, were the only monophyletic subgenera. Three major clades were strongly supported. That study underscored the need for additional molecular data to better resolve species level relationships, particularly for polyploids. Asian *Rubus* species were examined using limited nuclear and chloroplast loci by Wang et al. (2016). Species from *Dalibardastrum* and *Idaeobatus* were nested within the paraphyletic *Malachobatus*. These authors hypothesized that the allopolyploid species in *Malachobatus* may be derived from crosses between *Idaeobatus* and *Cylactis* species (Wang et al., 2015; Wang et al., 2016). *Idaeobatus* was polyphyletic with members in four clades. Current phylogenies consistently indicate that subgeneric labels rarely represent monophyletic groups (Alice and Campbell, 1999; Wang et al., 2016).

Hybridization and polyploidization are major evolutionary forces in *Rubus*. Intraspecific morphological and ploidal variability and the capability of many species to hybridize widely across the genus complicate traditional taxonomic classification (Bammi and Olmo, 1966; Alice et al., 2001; Mimura et al., 2014; Wang et al., 2015). Past phylogenetic analyses of the genus were based on nuclear ribosomal DNA internal transcribed spacer (ITS) sequence data and a few other nuclear and chloroplast loci, including *GBSSI-2, PEPC, trnL/F, rbcL, rpl20-rps*12, and *trnG-trnS* (Alice and Campbell, 1999; Yang and Pak, 2006; Wang et al., 2016). Relying on a limited number of loci to determine relationships in this genus with prevalent hybridization and polyploidy has resulted in low phylogenetic resolution. Additionally, single gene trees may not represent species trees due to hybridization, incomplete lineage sorting (ILS), and gene duplication (Maddison, 1997).

Two contrasting views of *Rubus* evolution exist. One view uses a nuclear ribosomal ITS-based genus-wide phylogeny (Alice and Campbell, 1999) to suggest that the ancestral area for the genus was North America, Eastern Europe (possibly Russia) or Asia (possibly Korea or Japan). In contrast, the treatment of Chinese *Rubus* by Lu (1983) hypothesizes that China, where *Rubus* is species-rich, is the origin of the genus.

In an analysis of Rosaceae using 19 fossils, 148 species and hundreds of low copy nuclear loci, Xiang et al. (2016) estimated that this genus originated in the Late Cretaceous approximately 75 million years ago (Ma). Zhang et al. (2017) estimated the age of the root node in a family-wide study of plastid sequences to be 57–66 Ma. *Rubus* fossils exist from the Tertiary period in the Eocene, which began ~55 Ma, and the more recent Oligocene, Miocene and Pliocene ages, on both sides of the North American land bridge and the Bering land bridge (Graham, 2018).

Certain biogeographical aspects are important to consider for *Rubus* evolution. The North American land bridge connected eastern North America with Europe and Asia before breaking up ~30 Ma, while the Bering land bridge remained intact until ~5 Ma (Tiffney, 1985; Milne, 2006). Both of these land bridges were important distribution avenues for subtropical (during the warmer Eocene) and temperate species throughout the Tertiary period (Tiffney, 1985; Wen, 1999; Wen, 2001; Wen et al., 2016). The Panamanian Isthmus connecting Central and South America began closing during the Paleogene approximately 30 Ma. It was crossable for plants and animals at approximately 20 Ma before finally closing 3 Ma (O'Dea et al., 2016).

Target capture allows hundreds to thousands of targeted loci to be sequenced for multiple individuals efficiently within a single high-throughput sequencing using Illumina® (San Diego, CA) lane. This technique has resolved phylogenetic questions across a

TABLE 1 | Accessions of *Rubus* species and outgroup (*Waldsteinia fragarioides*) used in this study.

Species	Ploidy	USDA GRIN subgenus classification	Focke subgenus classification	Region of origin	Group (1–8)	Voucher
R. deliciosus Torr.	2*x*	*Anoplobatus*	*Anoplobatus*	North America	2	PI 553184/CRUB 1021.001
R. odoratus L.	2*x*	*Anoplobatus*	*Anoplobatus*	North America	2	Alice R14, MAINE
R. parviflorus Nutt.	2*x*	*Anoplobatus*	*Anoplobatus*	North America	2	PI 553785/CRUB 13.001
R. trilobus Thunb.	2*x*	*Anoplobatus*	*Anoplobatus*	South America	2	Ruiz 889, MO
R. calycinus Wall. Ex D. Don	6*x*	*Chamaebatus*	*Chamaebatus*	Asia	5	Alice et al. (2008) Now Vouchered at WKU 04-07
R. nivalis Douglas	2*x*	*Chamaebatus*	*Chamaebatus*	North America	8	PI 679726/CRUB 1374.001 PL
R. pectinellus Maxim.*	6x	*Chamaebatus*	*Chamaebatus*	Asia	n/a	Jutila and Fujino 680, MO
R. pectinarioides	4*x*	*Chamaebatus**	n/a	Asia	5	Alice et al. (2008) Vouchered WKU 04-25
R. sengorensis	4*x*	*Chamaebatus**	n/a	Asia	5	Alice et al. (2008) Vouchered WKU 04-33
R. chamaemorus L.	8*x*	*Chamaemorus*	*Chamaemorus*	North America/ Northern Europe	1	Alice R17, MAINE
R. geoides Sm.	4*x*	*Comaropsis*	*Comaropsis*	South America	8	Dudley et al. 1538a, MO
R. arcticus L.	2*x*	*Cylactis*	*Cylactis*	North America/ Northern Europe	3	T. Eriksson 701, S
R. humulifolius C. A. Mey.	4*x*	*Cylactis*	*Cylactis*	Asia	4	PI 553242/CRUB 1173.001 PL
R. saxatilis L.	4*x*	*Cylactis*	*Cylactis*	Europe/Asia	7	PI 370230/CRUB 918.001 PL
R. lasiococcus A. Gray	2*x*	*Cylactis*	*Dalibarda*	North America	1	Merello et al. 827, MO
R. pedatus Sm.	2*x*	*Cylactis*	*Dalibarda*	North America/Asia	1	Alice 96-1, MAINE
R. fockeanus Kurz	4*x*	*Cylactis*	*Dalibarda*	Asia	5	PI 606537/CRUB 1960.000 SD
R. pubescens Raf.	2*x*	*Cylactis*		North America	3	Alice R15, MAINE
R. treutleri Hook. f.	4*x*	*Dalibardastrum*	*Dalibardastrum*	Asia	5	Alice et al. (2008) Now Vouchered at WKU 04-09
R. tricolor Focke	4*x*	*Dalibardastrum*	*Dalibardastrum*	Asia	5	Alice 97-2, MAINE
R. amphidasys Focke	6*x*	*Dalibardastrum*	*Malachobatus*	Asia	5	PI 618397/CRUB 1693.001 PL
R. nepalensis (Hook.f) Kuntze	4*x*	*Dalibardastrum*	n/a	Asia	5	Alice 97-1, MAINE
R. gunnianus Hook.	4*x*	*Diemenicus*	*Dalibarda*	Australia	8	Wells 96-1, MAINE
R. trifidus Thunb.	2*x*	*Idaeobatus*	*Anoplobatus*	Asia	4	PI 554051/CRUB 3.001 PL
R. parvifolius L.	2*x*	*Idaeobatus*		Asia	7	PI 553813/CRUB 5.001 PL
R. hawaiensis A. Gray	2*x*	*Idaeobatus*	*Idaeobatus*	North America (Hawaii)	3	PI 553214/CRUB 399.001 PL
R. spectabilis Pursh	2*x*	*Idaeobatus*	*Idaeobatus*	North America	3	PI 553980/CRUB 4.001 PL
R. crataegifolius Bunge	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	4	PI 553173/CRUB 16.001 PL
R. ellipticus Sm.	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	4	PI 553190/CRUB 1052.001 PL
R. illecebrosus Focke	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	4	PI 553643/CRUB 838.001 PL
R. palmatus Thunb.	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	4	PI 553782/CRUB 2.002 PL
R. rosifolius Sm.	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	4	Eurard 11660, MO
R. pentagonus Wall. Ex Focke	4*x*	*Idaeobatus*	*idaeobatus*	Asia	5	Alice et al. (2008) Vouchered WKU 04-06
R. thomsonii Focke	4*x*	*Idaeobatus*	*idaeobatus*	Asia	5	Alice et al. (2008) Vouchered WKU 04-31
R. alexeterius Focke	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	7	Alice et al. (2008) Vouchered WKU 04-23
R. coreanus Miq.	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	7	PI 618447/CRUB 1438.001 PL
R. idaeus L.	2*x*	*Idaeobatus*	*Idaeobatus*	Europe/Asia	7	T. Eriksson 735, S
R. innominatus S. Moore	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	7	PI 553646/CRUB 1039.001 PL
R. lasiostylus Focke	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	7	PI 553668/CRUB 425.001 PL
R. leucodermis Douglas ex Torr. & A. Gray	2*x*	*Idaeobatus*	*Idaeobatus*	North America	7	PI 553673/CRUB 14.001 PL
R. niveus Thunb.	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	7	PI 553723/CRUB 269.001 PL
R. occidentalis L.	2*x*	*Idaeobatus*	*Idaeobatus*	North America	7	AliceR16,MAINE
R. phoenicolasius Maxim.	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	7	Alice96-2,MAINE
R. pungens Cambess.	2*x*	*Idaeobatus*	*Idaeobatus*	Asia	7	PI 553849/CRUB 46.002 PL
R. sachalinensis H. Lév.	4*x*	*Idaeobatus*	*Idaeobatus*	Asia	7	PI 553866/CRUB 626.001 PL
R. strigosus Michx.	2*x*	*Idaeobatus*	*Idaeobatus*	North America	7	Maine Alice R8
R. macraei A. Gray	6*x*	*Idaeobatus*	*n/a*	North America (Hawaii)	6	Gardners. n., HPDL207
Logan	6*x*	*Idaeorubus*	*n/a*	Cultivar	7	PI 553258/CRUB 81.001 PL
Boysen	7*x*	*Idaeorubus*	*n/a*	Cultivar	8	PI 553341/CRUB 1108.001

(Continued)

TABLE 1 | Continued

Species	Ploidy	USDA GRIN subgenus classification	Focke subgenus classification	Region of origin	Group (1–8)	Voucher
Marion	6*x*	*Idaeorubus*	*n/a*	Cultivar	8	PI 553254/CRUB 385.001 PL
R. assamensis Focke	4*x*	*Malachobatus*	*Malachobatus*	Asia	5	PI 618433/CRUB 1701.001 PL
R. ichangensis Hemsl. & Kuntze	4*x*	*Malachobatus*	*Malachobatus*	Asia	5	PI 618453/CRUB 1606.001 PL
R. irenaeus Focke	6*x*	*Malachobatus*	*Malachobatus*	Asia	5	PI 618550/CRUB 1607.001 PL
R. lambertianus Ser.	4*x*	*Malachobatus*	*Malachobatus*	Asia	5	Boufford and Bartholomew 23955, MO
R. lineatus Reinw.	4*x*	*Malachobatus*	*Malachobatus*	Asia	5	Grierson and Long 1950, GH
R. clinocephalus Focke	4*x*	*Malachobatus*	*Malachobatus*	Asia	5	PI 606459/CRUB 1642.001 PL
R. tephrodes Hance	4*x*	*Malachobatus*	*Malachobatus*	Asia	5	Yao 9231, MO
R. australis G. Forst.	4*x*	*Micranthobatus*	*Lampobatus*	New Zealand	8	Gardner 1539, MO
R. parvus Buchanan	4*x*	*Micranthobatus*	*Lampobatus*	New Zealand	8	Alice 97-3, MAINE
R. moorei F. Muell.	4*x*	*Micranthobatus**	*Lampobatus*	Australia	8	Streimann 8207, GH
R. calophyllus	4*x*	*n/a*	*Malachobatus*	Asia	5	Alice et al. (2008) Vouchered WKU 04-24
R. repens (L.) Kuntze	2*x*	*n/a*	*Dalibarda*	North America	1	Alice 97-4, MAINE
R. ursinus × *R. armeniacus (1)*	8*x*	*n/a*	*n/a*	North America	8	Alice personal collection
R. ursinus × *R. armeniacus (6)*	8*x*	*n/a*	*n/a*	North America	8	Alice personal collection
R. acanthophyllus Focke	6*x*	*n/a*	*Orobatus*	South America	8	Alice and Cantrell are collectors in Ecuador WKU 07-11
W. fragarioides (Michx.) Tratt.	2*x*	*n/a*	*n/a*	North America	Outgroup	Hill & Soblo 21384, GH
R. glabratus Kunth	6*x*	*Orobatus*	*Orobatus*	South America	8	PI 548901/CRUB 1251.004 PL
R. loxensis Benth.	6*x*	*Orobatus*	*Orobatus*	South America	8	Alice and Cantrell are collectors in Ecuador WKU 07-17
R. roseus Poir.	6*x*	*Orobatus*	*Orobatus*	South America	8	Luteyn and Quezada 14402, MO
R. laegaardii Romol.	6*x*	*Orobatus**	n/a	South America	8	Voucher WKU 07-15
R. hispidus L. *	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	n/a	Alice R9, MAINE
R. caesius L.	4*x*	*Rubus* (= *Eubatus*)	*Eubatus*	Europe/Asia	6	Karlen 243, S
R. ursinus Cham. Et. Schltdl. *(2)*	8*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	6	PI 604641 CRUB 1857.001 PL
R. ursinus(3)	12*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	6	PI 554067/CRUB 197.001 PL
R. ursinus(4)	13*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	6	USDA Accession no longer exists
R. ursinus(5)	6*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	6	PI 604641/CRUB 1857.001 PL
R. allegheniensis Porter	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	8	Alice R1, MAINE
R. argutus Link	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	8	Alice & Judd 15, MAINE
R. armeniacus Focke	4*x*	*Rubus* (= *Eubatus*)	*Eubatus*	Europe/Asia	8	PI 618579/CRUB 45.001 PL
R. bifrons Vest	4*x*	*Rubus* (= *Eubatus*)	*Eubatus*	Europe/Asia	8	Alice 98-9, MAINE
R. canadensis L.	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	8	Alice & Campbell 98-10, MAINE
R. caucasicus Focke	4*x*	*Rubus* (= *Eubatus*)	*Eubatus*	Europe/Asia	8	PI 553143/CRUB 54.001 PL
R. coriifolius Liebm.	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	Americas	8	Vouchered WKU 06-05
R. cuneifolius Pursh	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	8	Alice 5, MAINE
R. flagellaris Willd.	4-9*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	8	PI 553787/CRUB 61.001 PL
R. laciniatus Willd.	4*x*	*Rubus* (= *Eubatus*)	*Eubatus*	Europe/Asia	8	PI 618548/CRUB 1596.001 PL
R. robustus C. Presl	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	Americas	8	Steinbach 247, GH
R. setosus Bigelow	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	8	Alice 113, MAINE
R. trivialis Michx.	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	North America	8	Alice 33, MAINE
R. ulmifolius Schott	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	Europe/Asia	8	190-84, MOR
R. urticifolius Poir.	2*x*	*Rubus* (= *Eubatus*)	*Eubatus*	Americas	8	PI 548929/CRUB 1288.001 PL
R. glaucus Benth.	4*x*	*Rubus* (= *Eubatus*)	*Idaeobatus*	South America	6	PI 548906/CRUB 1293.001 PL
R. eriocarpus Liebm.	2*x*	*Rubus* (= *Eubatus*)	*Idaeobatus*	South America	7	Vouchered WKU 06-12
R. pensilvanicus Poir.	4*x*	*Rubus* (= *Eubatus*)	n/a	North America	8	Alice R5, MAINE

Species marked with an asterisk in the "Species" column did not sequence well and were not included in the results. Subgenera classifications in Focke and the USDA GRIN network are reported. Subgenera marked with an asterisk in the "USDA GRIN Subgenus Classification" column are not listed in GRIN. Focke subg. Eubatus *has been renamed to subg.* Rubus*. Current classifications were curated from other publications (Barneby, 1988; Bean, 1995; Romoleroux et al., 1996; Sutherland, 2005). Herbarium vouchers with collector, number, and herbarium (Holmgren et al., 1990) or PI numbers for accessions of plants housed in the living collection at USDA NCGR Corvallis are given. MOR refers to the living collection at Morton Arboretum, Lisle, IL. HPDL refers to the Native Hawaiian Plants DNA library (Morden et al., 1996). The geographic origin for each accession is listed by continent or region. Ploidy data was collected from flow cytometry data, multiple publications, and the Missouri Botanical Garden index of plant chromosome number database (Thompson, 1995; Thompson, 1997; Meng and Finn, 2002; Hummer et al., 2015). Eight major phylogenetic groups were identified in nuclear sequence analyses. The group in which each species is found is listed.*

range of plant genera, including *Asclepias* L. (Weitemier et al., 2014), *Heuchera* L. (Folk et al., 2017), and *Lachemilla* L. (Morales-Briones et al., 2018). Although not specifically targeted, chloroplast sequences can be obtained after sequencing target capture libraries, enabling an independent estimate of phylogeny and inference from a predominantly maternally inherited genome (Weitemier et al., 2014; Folk et al., 2017; Dillenberger et al., 2018).

Our objectives were to estimate phylogenetic relationships in *Rubus* using a large molecular dataset over a genus-wide species sampling; estimate divergence times between major *Rubus* clades; and examine the biogeography of species diversification.

MATERIALS AND METHODS

Samples

Samples designated by a plant information (PI) number (**Table 1**), were obtained from the US Department of Agriculture (USDA ARS NCGR) according to rules of the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGR, 2019). DNA from leaf samples without PI numbers were obtained by LA through field work, and exchange from international botanical gardens and herbaria (**Table 1**).

Sampling and DNA Extraction

We sampled 94 accessions, representing 87 wild *Rubus*, three cultivars (*R.* hybrid "Logan," "Boysen," "Marion"), and outgroup *Waldsteinia fragarioides* (**Table 1**). *Rubus* is sister to the clade containing *Waldsteinia* in the phylogeny of Rosaceae estimated by Potter et al. (2007) and Xiang et al. (2016). Twenty-six species are from subg. *Idaeobatus*, 24 are from subg. *Rubus* and other subgenera are represented by 1–9 species each (**Supplementary Table S1**).

"Logan," "Boysen," and "Marion" were sampled because they are economically important hybrid cultivars with known percentages of blackberry and raspberry parentage. "Logan" is comprised of 50% blackberry/50% raspberry species; "Boysen," an offspring of "Logan," is 75% blackberry/25% raspberry; and "Marion" is 69% blackberry/31% raspberry (Jennings, 1988; Thompson, 1997).

Genomic DNA was isolated from fresh leaves frozen at −80°C, leaves dried in silica gel desiccant, or herbarium specimens (Holmgren et al., 1990; Morden et al., 1996; Alice and Campbell, 1999; Alice et al., 2008) using a modified CTAB (hexadecyltrimethylammonium bromide) extraction method (Doyle and Doyle, 1987).

Target Enrichment Probe Design

Targets were developed from within the genus or from closely related genera within Rosaceae. We used the *Rubus occidentalis* genome v1 assembly (VanBuren et al., 2016) and a conserved set of loci from *Fragaria vesca*, *Malus* × *domestica* and *Prunus persica* (Liston, 2014). Exon sequences were extracted from the *R. occidentalis* transcriptome assembly (VanBuren et al., 2016). Only those exons ≥ 80 bp, with GC content between 30 and 70%, and with one BLAST hit to the *R. occidentalis* genome over 50% of the exon length and with ≥ 90% identity were used for bait development. In total, probes were synthesized by MYcroarray (now Arbor Biosciences, Ann Arbor, MI, USA) for 8,963 exons from 926 genes. Due to a bioinformatics error, the *R. occidentalis* exon sequences from which probes were created were cropped into 60 bp sequences separated by 20 bp gaps before submission to MYcroarray. The 120-mer baits synthesized by MYcroarray with 1x tiling corresponded to 140 bp of genome sequence. Despite this, hybridization with the *R. occidentalis*-derived probes was successful for nearly all study samples.

Conserved loci from *F. vesca*, *Malus* × *domestica* and *P. persica* genomes were selected for their usefulness in comparative genomic studies across Rosaceae as described by Liston (2014). Briefly, single copy loci shared between the *F. vesca* and *P. persica* genomes were identified. The corresponding genes were extracted from the *Malus* × *domestica* genome, where there were often two gene copies due to the allopolyploid ancestry of the former Rosaceae subfamily Maloideae. The gene sequence with the fewest ambiguous bases or polymorphic sites was selected. Genes were filtered based on their phylogenetic utility (≥ 960 bp, > 85% pairwise sequence similarity between the three genomes) and to maximize the success of target capture (exons ≥ 80 bp, GC content > 30% or < 70%, < 90% sequence similarity to other target exons in the same genome). This resulted in 257 genes; probes were designed for the copies of these genes originating from *F. vesca*.

Library Preparation

Genomic DNA was quantified with PicoGreen (ThermoFisher Scientific, Waltham, MA, USA) and quality checked using agarose gel electrophoresis. To prepare for library construction, 400 ng of input DNA was sonicated for 5–10 min using a Diagenode BioRuptor Sonicator (Denville, NJ, USA). After an initial 5 min of sonication, samples were sized using gel electrophoresis and sonicated an additional 1–5 min as necessary to achieve the desired 200 bp average insert size. If DNA bands were very faint after the first round of sonication, a new aliquot of the sample with 600–800 ng of input DNA was prepared and sonicated. Sonicated samples were cleaned using Qiaquick PCR purification columns (QIAGEN, Valencia, CA, USA) to eliminate low molecular weight fragments. Genomic libraries were prepared using the NEBNext Ultra DNA Library Prep Kit with NEBNext Multiplex Oligos for Illumina (New England Biolabs, Ipswich, MA, USA) to enable multiplexed sequencing. Size selection for 200 bp fragments was done after adaptor ligation using AMPure (Agencourt Bioscience Corporation, A Beckman Coulter Company, Beverly, MA, USA) beads at a 0.55:1 ratio with the sample. Libraries were amplified for 8 PCR cycles and cleaned with AMPure beads at a 1:1 ratio with the sample before being quantified with PicoGreen. A subset of libraries was quality checked with the Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) at Oregon State University's (OSU) Center for Genome Research and Biocomputing (CGRB).

To prepare for in-solution hybridization, samples were divided into four pools of 24 samples containing 20 ng of each library. MYcroarray MYbaits (Arbor Biosciences, Ann Arbor, USA) protocol version 1.3.8 was followed for sequence enrichment. The resulting pools were quantified using Qubit and qPCR, pooled again in equimolar amounts and sequenced with 100 bp paired end reads in one Illumina® HiSeq ™ 2000 lane at the CGRB. Libraries were demultiplexed using the Illumina pipeline.

Sequence Assembly

Bases with a quality score under Q20 were trimmed from the right and left side of reads with BBduk; reads shorter than 25 bp after trimming were discarded (Bushnell, 2014). Adapters were not trimmed from reads, however very few adapter sequences were present in the read pool after quality trimming and therefore likely had a negligent impact on downstream analyses. When reads were checked for adapters using BBduk, no reads were discarded from the read pool and 99.78% of the bases were non-adapter sequence. Loci were assembled with HybPiper v. 1.2 using sequence read files and a target sequence reference file from which probes were designed (Johnson et al., 2016). To replace the missing 20 bp sequences from the *Rubus* baits in this target reference file, the 60 bp target fragments used in probe synthesis were first mapped against the *R. occidentalis* genome with BBMap. Then, Bedtools v. 2.25.0 was used to extract contiguous sequences for each exon (Bushnell, 2014; Quinlan, 2014). Exons for each gene were then concatenated to create the final target sequence reference. HybPiper creates bins based on reads by target sequence using BWA (Li and Durbin, 2009). The reads are then assembled with SPAdes into contigs using the target sequence as a reference (Bankevich et al., 2012; Li, 2013). Output sequences were either assembled exons or supercontigs, which could include noncoding sequences such as introns, 5′ UTR, and 3′ UTR sequences obtained from genomic libraries during hybridization.

Exons and supercontig sequences were each aligned with MAFFT v. 7.402. Alignment sites with gaps in more than 20% of sequences were removed with TrimAl v. 1.2rev59 to prevent ambiguous placement of taxa in a tree due to insufficient phylogenetic signal (Capella-Gutiérrez et al., 2009). Alignments were visually inspected for quality and removed if necessary. This resulted in 941 genes used in downstream analyses.

Phylogenetic Analyses of Nuclear Loci

The maximum likelihood phylogeny was estimated twice for each locus, once with the exon sequences and secondly with the supercontig sequence data. RAxML v. 8.1.21 was used to conduct a bootstrap search with up to 1000 replicates (-#autoMRE or -#1000 option) and estimate the maximum likelihood phylogeny for each gene [option –f a; Stamatakis (2014)]. The best fit model of evolution (GTRGAMMA or GTRGAMMAI) was determined with PartitionFinder v. 2.1.1 for the exon sequences of each gene. This same model was also used for supercontig sequence analyses (Lanfear et al., 2012). Phylogenies were estimated for two sets of taxa: one containing only diploids and the other containing all taxa polyploids and diploids. Thus, for each gene, a phylogeny was estimated for the following datasets: diploid exons, diploid supercontig sequences, all taxa exons, and all taxa supercontig sequences.

To prevent ambiguous placement of taxa in a tree resulting from insufficient phylogenetic signal, RogueNaRok v. 1.0 was used with default settings to identify such "rogue" taxa for each locus using bootstrapped RAxML trees (Aberer et al., 2013). Wilkinson and Crotti (2017) argued that this technique may be poorly suited to detecting rogue taxa, however, the automated reproducible approach RogueNaRok was chosen because this application simultaneously evaluated 941 gene trees. This large gene number supports an automated approach (Borowiec, 2019). Rogue taxa were eliminated from sequence alignments and gene trees were re-estimated with RAxML.

Species phylogenies were estimated under the multi-species coalescent model using ASTRAL-II v. 4.10.12 and SVDQuartets implemented in PAUP* 4.0 (Swofford, 2003; Chifman and Kubatko, 2014; Mirarab and Warnow, 2015). ASTRAL-II and SVDQuartets both use relationships between quartets of taxa to estimate the overall species tree. ASTRAL-II identified the species tree that shares the maximum number of quartet trees with the 941 gene trees estimated with RAxML (Mirarab and Warnow, 2015). Local posterior probability support values were calculated as these have been shown to be highly precise compared with multi-locus bootstrapping (Sayyari and Mirarab, 2016). SVDQuartets randomly sampled 100,000 possible quartets of taxa and used SNPs from the concatenated sequence alignments to score each possible split in the quartets [100 bootstrap replicates; (78)]. The best scoring splits were assembled into a species phylogeny in PAUP* using QFM (Swofford, 2003; Reaz et al., 2014).

Branch support for phylogenies with the highest likelihood for each concatenated sequence alignment were also evaluated using Quartet Sampling (Pease et al., 2018). This method evaluates the topological relationship between quartets of taxa using an input phylogeny and a molecular alignment partitioned by gene. Unlike bootstrap values, this method can distinguish if the data supporting internal branches is strongly discordant or lacking signal (Pease et al., 2018). Quartet Sampling produces three main scores, quartet concordance (QC), quartet differential (QD), and quartet informativeness (QI) for each node. Quartet concordance describes how often concordant quartets, which show the same splits and sister relationships between clades, are inferred. Scores ≥ 0.5 indicate strong support for the concordant topology. Quartet differential measures how often quartets with discordant topologies are inferred. This measure can indicate if a dataset shows strong evidence for an alternate evolutionary history at a node. Scores ~1 indicate that no alternate topology is strongly favored. Quartet informativeness measures the proportion of replicates that are informative for a node. Scores = 1 indicate that all replicates were informative while scores = 0 indicate that none were informative.

Network Analysis

Because we anticipated high levels of ILS and hybridization in this dataset, unrooted super networks were estimated to visualize incongruences among exon or supercontig sequence gene trees

and identify putative hybrid taxa using SuperQ v. 1.1 with the Gurobi optimizer and a balanced linear secondary objective function (Grunewald et al., 2013). In this method, input gene trees (identical to gene trees used in ASTRAL-II analyses) are broken down into quartets and reassembled into a network where edge lengths indicate the frequency of each split in the gene tree set.

Dating for Phylogenetic Estimation

ASTRAL-II-generated topologies from genes estimated using exon sequences were used for dating. Branch lengths per site substitution rates were estimated over the ASTRAL-II topology for all taxa using RAxML [-f e option, GTRGAMMA model of evolution; Stamatakis (2014)] and the corresponding concatenated alignment of exon sequences. Phylogenies were dated with r8s version 1.80 using the penalized likelihood method and the truncated Newton algorithm with a smoothing parameter estimated using cross validation (Sanderson, 2002; Sanderson, 2003). The age of the root node was constrained to 56.93–65.66 Ma based on the age of this node estimated from plastid sequences (Zhang et al., 2017).

Biogeographic Analyses

Data were collected for the continent of origin for each sample (**Table 1**). Ancestral ranges were estimated with BioGeoBEARS version 1.1 over ultrametric dated phylogenies resulting from r8s using Dispersal-Extinction-Cladogenesis (DEC) and DEC+*j* likelihood models (Ree and Smith, 2008; Matzke, 2014).The parameter *j* incorporates founder-event speciation or long distance dispersal events (Matzke, 2013; Matzke, 2014). The DEC+j had the lowest AIC but it's controversial to compare the DEC+J and DEC models with this metric (Andersen et al., 2018; Lu et al., 2018; Leavitt et al, 2018). The DEC model results have the lowest AIC value compared with the DIVALIKE and BAYAREALIKE models so the DEC tree is presented (**Figure 4**).

Chloroplast Sequence Extraction and Analysis

Reads for each sample were mapped to the *R. occidentalis* chloroplast reference genome (VanBuren et al., 2016) edited with BBMap to contain only one copy of the inverted repeat (Bushnell, 2014; VanBuren et al., 2016). Consensus chloroplast sequences from a reduced read set of up to 100,000 mapped reads were extracted using Geneious v. 9.1.7 with Ns inserted at sites with no sequence coverage (Kearse et al., 2012). Consensus sequences were aligned with MAFFT using auto settings (Katoh and Standley, 2013). Alignment sites with missing data in over 20% of samples were stripped using Geneious v. 9.1.7 (Kearse et al., 2012). The maximum likelihood phylogeny was estimated with RAxML using up to 1000 bootstrap replicates Stamatakis (2014) under the GTRGAMMAI model of evolution. Rogue taxa were identified with RogueNaRok and removed from the alignment (Aberer et al., 2013). RAxML was subsequently run to estimate the final maximum likelihood phylogeny.

RESULTS

Sequencing Target Genes and Chloroplast Genome

The average sequencing depth for all samples over all loci was 66.8x (**Supplementary Table S2**). The samples of *Rubus hispidus* and *R. pectinellus* had an average sequencing depth across all loci under 1x and contigs for <10 genes and were excluded from phylogenetic analyses. The average percentage of on-target reads was 71.3%. HybPiper produced sequences for an average of 1,113 genes per taxon, and an average of 988 sequences were at least 75% of the target length. An average of 86% of target bases were recovered for genes shared across Rosaceae and 101% of bases for *R. occidentalis* targets (**Supplementary Table S2**). Alignment lengths for supercontigs, i.e., exons + noncoding sequences, were 10.1 Mbp for diploid species only (average ungapped length 3.8 Mbp) and 10.1 Mbp for polyploid and diploid taxa (average ungapped length 2.7 Mbp) (**Supplementary Table S3**). The concatenated alignment length of exon sequences for each gene was 2.5 Mbp for diploid species only (average ungapped length 1.6 Mbp) and 2.5 Mbp for all analyzed taxa (average ungapped length 1.7 Mbp). The supercontig sequence alignments of diploids and all species had 17% and 23% variable sites and 7% and 11% phylogenetically informative sites, respectively. Exon alignments were 20% variable (9% phylogenetically informative) for diploids and 29% variable (15% informative) for all species analyzed.

After automated trimming and manual evaluation of alignment quality, 941 gene targets remained for exon alignments and 905 to 910 for supercontigs from all taxa, and from diploids only, respectively (Carter, 2018). After removal of rogue taxa (those with ambiguous phylogenetic placement), exon alignments of all taxa and alignments of only diploid taxa contained an average of 52 (55% of total sample set) and 30 taxa (70% of total sample set), respectively. Supercontig alignments including all taxa contained an average of 39 taxa (41% of total sample set), while alignments of only diploid taxa contained 27 individuals on average, or 63% of total sample set (Carter, 2018).

The chloroplast alignment of sequences from 89 taxa was 125,795 bp. RogueNaRok identified *R. caucasicus, R. lambertianus* and *R. robustus* as rogue taxa and they were removed from the chloroplast analysis. Average coverage of the 127,679 bp *R. occidentalis* reference genome was 24x, ranging from 1.3x–99.6x (**Supplementary Table S4**).

Phylogenetic Analyses

Differences between the ASTRAL-II and SVDQuartets analyses for all taxa and diploid-only taxa datasets were more evident in the topology of internal nodes delineating the relationships between groups (**Figure 1**). These nodes represent relationships between groups that may commonly hybridize or where ancestors of extant taxa may have been progenitors of multiple clades. Deep evolutionary signal for these events may have been obscured by more recent polyploidization and

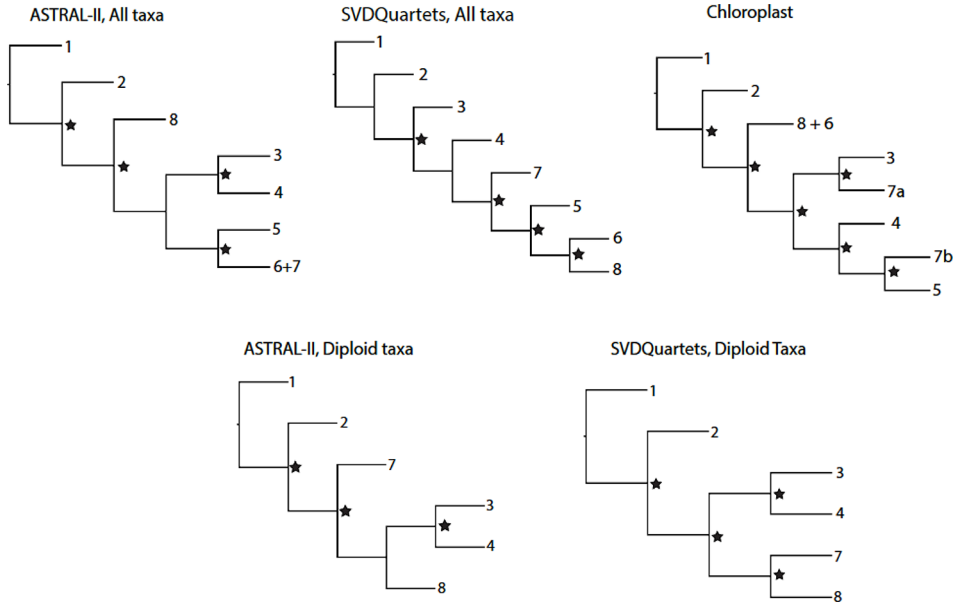

FIGURE 1 | Topological relationships between genus *Rubus* Groups 1–8 in phylogenetic analyses of exon or chloroplast sequences. Nodes with strong support (Bootstrap > 75 for SVDQuartets phylogenies; Posterior Probability > 0.95 for ASTRAL-II phylogenies) are marked with a star.

hybridization events, leading to topological conflict between analyses. The quartet concordance (QC) values for two nodes describing relationships between major groups in the SVDQuartets phylogenies indicate counter support for the topology. The alternate topologies seen in the ASTRAL-II trees have weak support and skewed distributions for discordant topology frequencies at some internal nodes (Carter, 2018). The SVDQuartets trees are likely exhibiting these discordant topologies that are supported by a significant minority of loci. In a previous report, ASTRAL-II phylogenies were shown to be more accurate than SVDQuartets trees in the presence of high ILS (Chou et al., 2015).

The supercontig sequence alignments contained a high proportion of missing data. On average, 73% of the data was missing from the supercontig sequence alignments for all taxa, compared to 42% of missing data for the exon alignment for all taxa (Carter, 2018). Similarly, diploid alignments had an average of 64% missing data for supercontig sequence data and 39% for exon sequences. When compared, the exon-only phylogenies and the supercontig sequence trees show the same major groups of taxa and similar variations in backbone topologies between analyses (**Figure 1**; exon-only phylogenies). Because the supercontig sequence dataset did not provide additional phylogenetic resolution and contained less complete alignments, the exon sequences were analyzed.

Eight consistent groups of taxa corresponding roughly to eight clades were seen in the SVDQuartets and ASTRAL-II generated phylogenies from all datasets: diploid exons, diploid supercontig sequences, polyploid and diploid exons, and polyploid and diploid supercontig sequences (**Figures 2** and **3**). Most relationships in the analyses were well-supported (bootstrap values > 75; posterior probabilities > 0.95). In addition, group 8 was divided into 8a representing the majority of this clade and subg. *Rubus*; group 8b, including subg. *Orobatus* species; and group 8c, including subg. *Comaropsis, Diemenicus,* and *Micranthobatus*. Most relationships in the analyses were well-supported (bootstrap values > 75; posterior probabilities > 0.95).

Groups 1 and 2 include eight species from subg. *Chamaemorus*, *Dalibarda*, *Cylactis*, and *Anoplobatus* and are sister to the remainder of genus *Rubus* (**Figure 2**, **Table 1**). Group 3 includes *R. hawaiensis* (*Idaeobatus*), *R. spectabilis* (*Idaeobatus*), *R. pubescens* (*Cylactis*), and *R. arcticus* (*Cylactis*), and is monophyletic. Group 4 is sister to Group 3 and contains seven taxa; six are classified in *Idaeobatus* and one in *Cylactis* (*R. humulifolius*). Group 5 consists of accessions of Asian origins from *Malachobatus*, *Daliardastrum*, *Cylactis*, *Idaeobatus*, and *Chamaebatus*. It is often sister to Group 7, which contains primarily *Idaeobatus* species with one *Cylactis* accession (*R. saxatilis*) and "Logan," a hybrid cultivar. Group 6, contains four of six *R. ursinus* accessions, *R. caesius*, and *R. glaucus* from subg. *Rubus* and *R. macraei* from *Idaeobatus,* and shifts positions between analyses but groups with either Group 7 or 8. Group 8 contains the most species and consists of accessions from subg. *Rubus* (8a), *Orobatus* (Group 8b), *Comaropsis*, *Micranthobatus*, and *Diemenicus* (8c), and the predominantly blackberry hybrid cultivars "Boysen" (75% blackberry/25% raspberry) and "Marion" (69% blackberry and 31% raspberry).

Anoplobatus and *Orobatus* are monophyletic (**Figure 2**). All other subgenera, except monotypic *Chamaemorus, Comaropsis* and *Diemenicus*, are para- or polyphyletic. *Anoplobatus* species comprise Group 2 and are sister to the majority of genus *Rubus*. *Orobatus* species form a subclade in Group 8 and are sister to the major subg. *Rubus* clade. Species from *Comaropsis,*

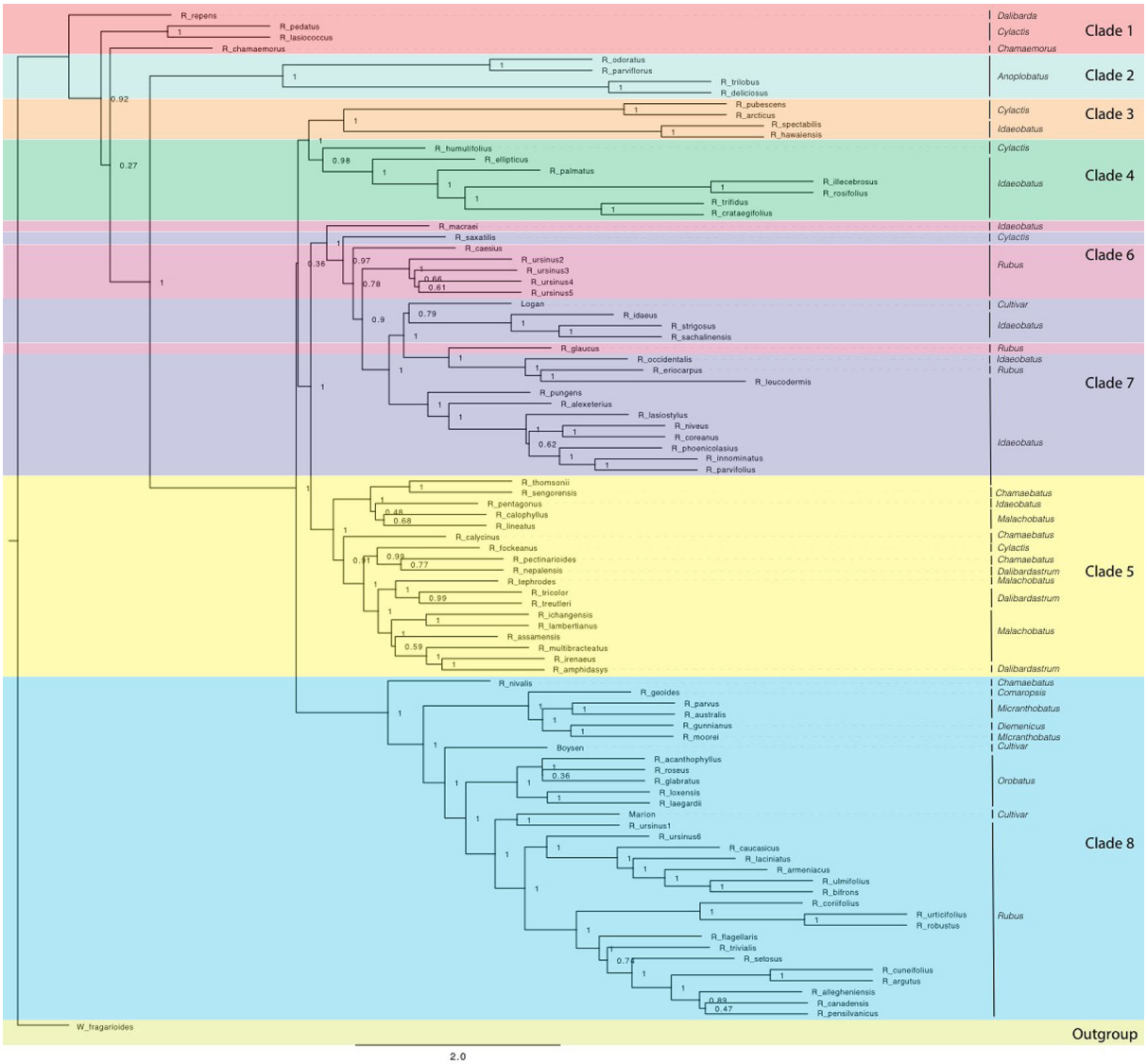

FIGURE 2 | ASTRAL-II phylogeny estimated from exon sequence gene trees from all *Rubus* taxa. Posterior probability values (0–1) are shown to the right of each node. Branch lengths are in coalescent units and measure discordance in the underlying gene trees. Groups are labelled with colored bands. Taxa are labelled with their subgeneric classification.

Micranthobatus, and *Diemenicus* also form a subclade in Group 8. Subg. *Rubus* would be monophyletic in Group 8 if not for *R. ursinus, R. glaucus,* and *R. caesius* in the variable Group 6, and *R. eriocarpus* in Group 7. These species are putative allopolyploids and are discussed below. Species from *Comaropsis, Micranthobatus,* and *Diemenicus* form a subclade in Group 8.

In the chloroplast phylogeny, Group 7 divides into two monophyletic clades. One is sister to Group 3 and the other to Group 5. The eight major groups also appear in phylogenetic network analyses (**Figures 4** and **5**).

Network analyses allowed a more thorough visualization of conflict within our data, particularly caused by hybridization, as discussed below, which cannot be captured in a dichotomously branching tree (**Figure 3**). Few of the assembled genes (0.47%, or 5.5 loci on average per taxon)) had paralogs. On average, polyploid taxa had paralogs in 8.2 loci compared to 2.4 loci for diploid taxa. Identification of paralogs by HybPiper is consistent with expectation that polyploids, with multiple subgenomes, would have a higher number of paralags than diploids (Veitia, 2005) (**Supplementary Table S5**).

Maternal and paternal progenitors of putative hybrid groups or species were assessed by comparing nuclear and chloroplast phylogenies (**Figures 2** and **3**). *R. nepalensis* and *R. allegheniensis* had long branch lengths compared to other taxa (**Figure 3**), likely due to limited sequence data for these samples (Carter, 2018). These species have sequences over 75% of the target

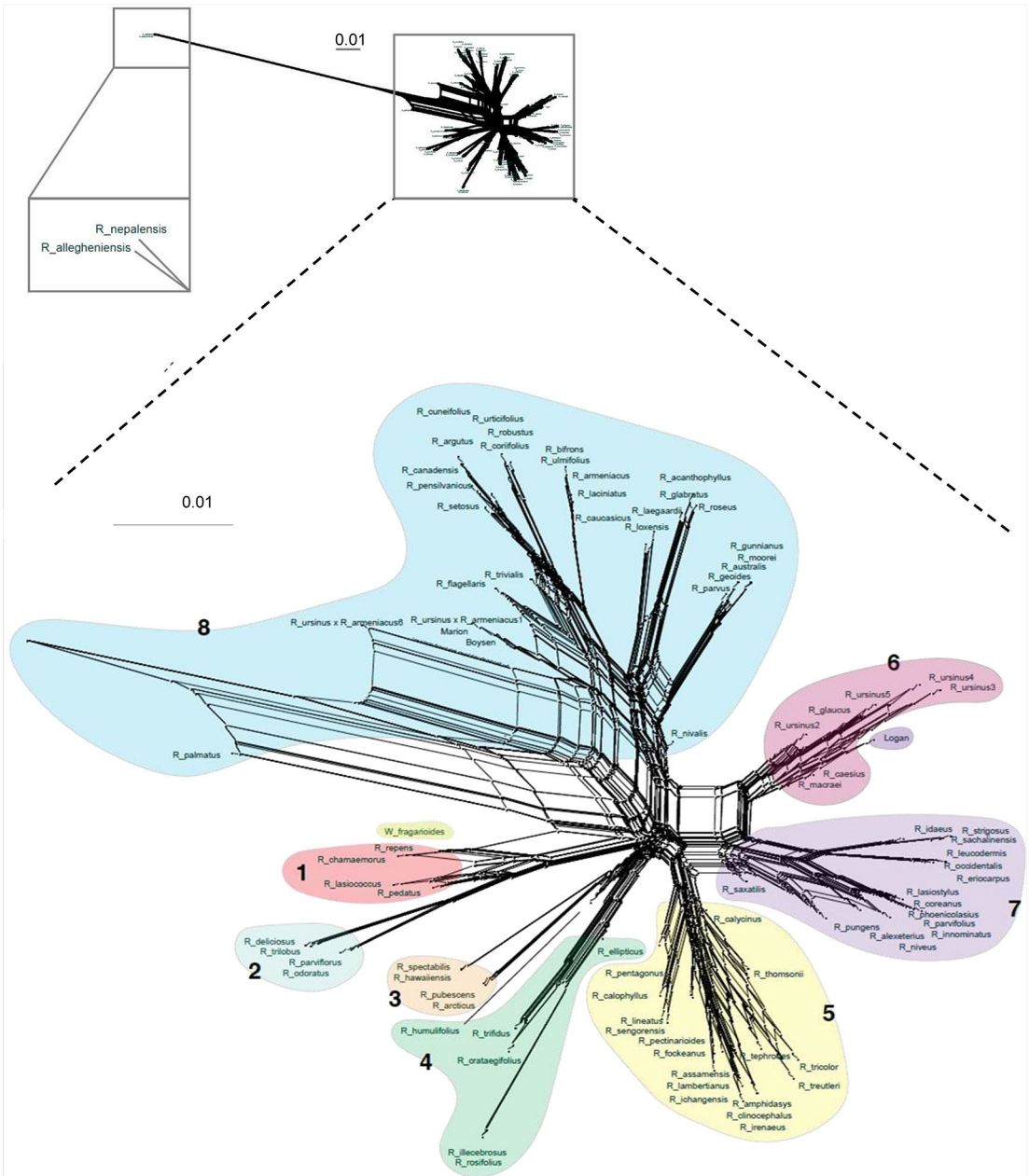

FIGURE 3 | Super network for all *Rubus* taxa estimated with SuperQ from exon gene trees estimated with RAxML. Colored shapes correspond to Groups 1–8. Top inset placement of *R. allegheniensis* and *R. nepalensis* due to limited sequence data for these samples (38).

length for only 89 and 66 targets, respectively. ASTRAL-II and SVDQuartets are robust to this level of missing data and place these species with high support in most species trees (**Figure 2**).

Phylogenetic Dating and Ancestral Range Estimation

Ultrametric trees of all taxa estimated from exon sequences and dated using r8s are shown (**Figure 4**). *Rubus* radiated throughout the Miocene with the eight major groups arising approximately 10–20 Ma. The DEC model for ancestral range estimation was rejected based on a likelihood ratio test ($p < 0.05$) and AIC values (Carter, 2018). Under the DEC+*j* model, the most likely ancestral range for *Rubus* for all taxa phylogenies was North America (**Figure 4**). Most recent common ancestors (MRCA) of Groups 1, 2, 3, and 8 were also most likely distributed in North America. Ancestral ranges in North America and Asia were similarly likely for Group 6 and 7 (**Figure 4**). Ancestors of Groups 4 and 5 were most likely distributed in Asia (**Supplementary Table S6**).

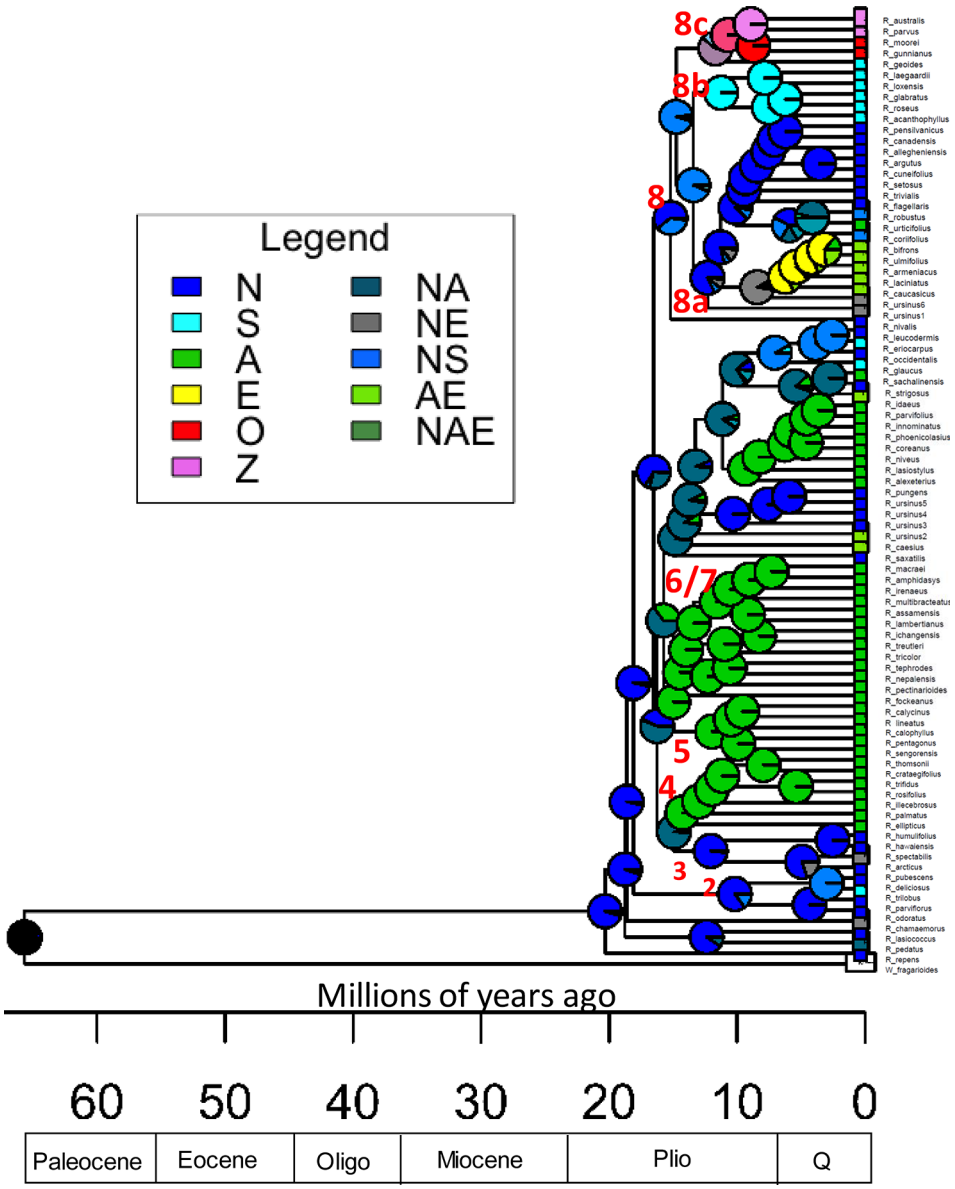

FIGURE 4 | *Rubus* ancestral range estimation using the DEC model for all taxa. Time scale is in millions of years. Pie charts represent relative probability of each area being the ancestral range. P, Pliocene; Q, Quaternary; N, North America (including Mexico and Guatemala); S, South America; A, Asia; E, Europe; O, Australia; Z, New Zealand. Combinations of letters indicate presence across multiple areas. Ancestral nodes for major groups are labelled numerically.

DISCUSSION

Phylogenetic Analyses and Taxonomic Implications

Our target capture sequencing approach enabled resolution of relationships between major groups, confirming or extending previous studies (Alice and Campbell, 1999; Morden et al., 2003; Yang and Pak, 2006; Wang et al., 2016). Subg. *Idaeobatus* is polyphyletic, as seen in studies of Asian and worldwide *Rubus* species (Alice and Campbell, 1999; Morden et al., 2003; Yang and Pak, 2006; Wang et al., 2016). *Rubus macraei* and *R. hawaiensis* have distinct evolutionary histories and likely resulted from separate colonization events of the Hawaiian Islands (Morden et al., 2003). *Rubus ursinus* is not closely related to *R. macraei* (**Figure 3**).

Rubus repens is placed in genus *Rubus* as *R. dalibarda* (Focke, 1910; Focke, 1911; Focke, 1914), but classified by other botanists in the monotypic genus *Dalibarda* due to unique morphological features rarely or not otherwise seen in *Rubus,* including dry fruits, reduced carpel number, and apetalous, carpellate and cleistogamous flowers (Bailey, 1941; Gleason and Cronquist, 1991; Alice and Campbell, 1999). Alice and Campbell (1999) showed this species nesting within *Rubus* using ITS and chloroplast data, spurring its reclassification into genus *Rubus*

(Alice et al., 2015). In our study, *R. repens* nests either within Group 1 or is sister to other *Rubus* species studied, supporting its classification in the genus *Rubus.*

The six subg. *Cylactis* species are distributed in Groups 1, 3, 4, 5, and 7 and often closely related to species in subg. *Chamaebatus* or *Idaeobatus* (**Figure 2**). Morphological differences used for current taxonomic classifications in Group 5 do not reflect phylogenetic relationships. Since the higher polyploids of this group may be allopolyploids with similar progenitor species, taxonomy based on morphology may be unreliable for this group.

Subgenus *Micranthobatus* is closely related to the monotypic subg. *Diemenicus* and *Comaropsis.* All species with known ploidy in these subgenera are tetraploid with small genomes (Hummer and Alice, 2017). Our results support the hypothesis of Hummer and Alice (2017) that these species may have descended from one allopolyploid ancestor, possibly a hybrid between diploids with small genomes. *Rubus nivalis,* a closely related diploid species, may have been a progenitor of this group. The common ancestor of these five species may have migrated from South America to the South Pacific through long distance dispersal by birds. Geographic isolation, potentially between populations of the common ancestor of *R. moorei* and *R. gunnianus*, may have led to strong morphological divergence. *Rubus gunnianus* of the monotypic subg. *Diemenicus* has unique morphological features, including leaves arising in rosettes directly from the rhizome, a lack of stipules, broad petioles, prominent carpel glands, and unisexual flowers (Bean, 1997).

Subg. *Rubus* species are primarily in Groups 6 and 8, with *R. eriocarpus* in Group 7. *Rubus eriocarpus* is morphologically similar to *R. glaucus* (Pankhurst, 2001). Both share stem and leaf characteristics with black raspberries but have fruit that retains the torus when picked (Standley and Steyermark, 1946; Jennings, 1988). *Rubus eriocarpus* is closely related to North American black raspberries *R. occidentalis* and *R. leucodermis* in nuclear and chloroplast phylogenies (**Figures 2** and **5**) while *R. glaucus* aligns with other putative blackberry × raspberry hybrids

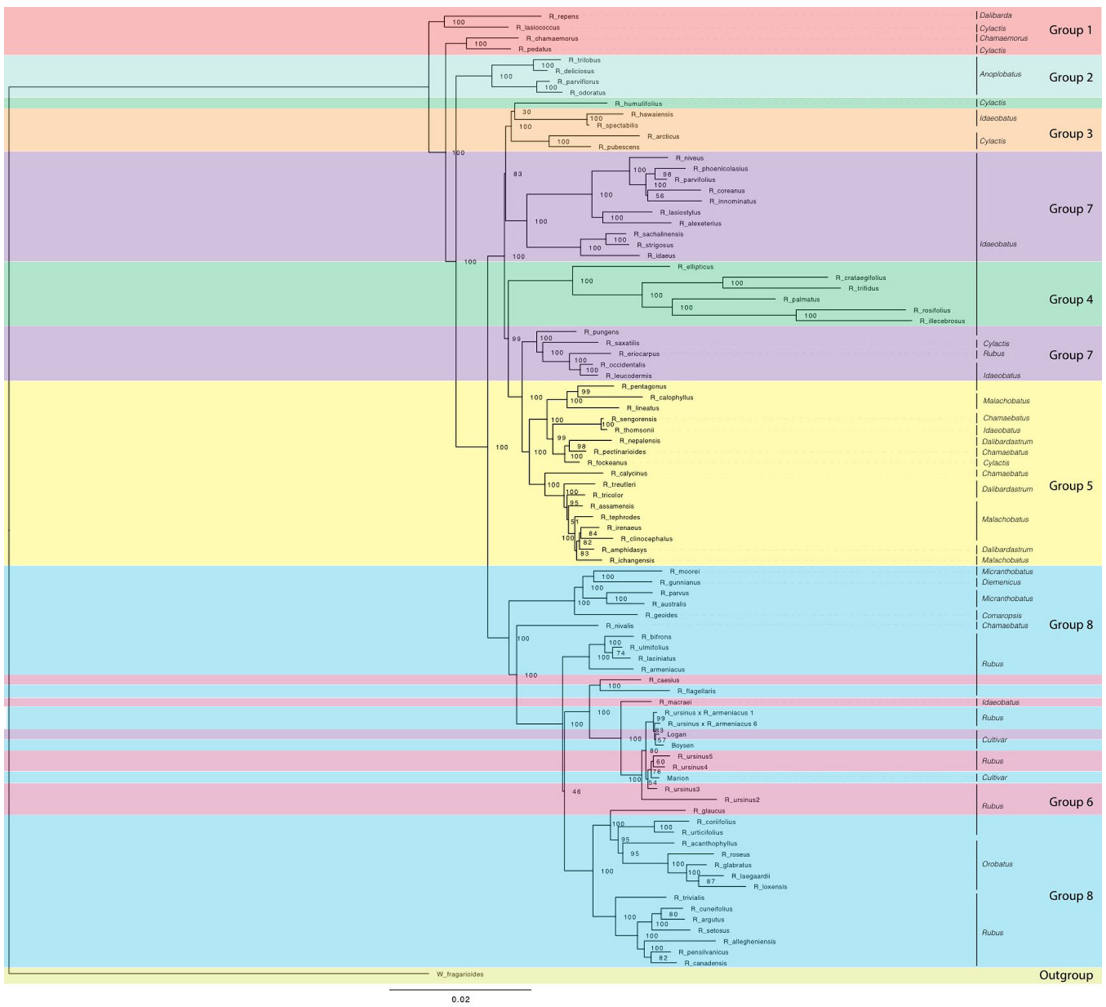

FIGURE 5 | Maximum likelihood phylogeny estimated with RAxML for chloroplast sequences from all *Rubus* taxa. Bootstrap values (0-100) are shown to the right of each node. Branch lengths represent relative evolutionary change. Groups are labelled with colored bands. Taxa are labelled with their subgeneric classification sensu GRIN (2019).

in Group 6. Focke (Focke, 1910; Focke, 1911; Focke, 1914) originally classified *R. eriocarpus* in *Idaeobatus*; our results support Focke's treatment of *R. eriocarpus* within subg. *Idaeobatus*. Similarities between *R. glaucus* and *R. eriocarpus* could be due to convergent evolution, or *R. eriocarpus* could be a parent of *R. glaucus*.

Subg. *Idaeobatus* is polyphyletic with representatives in Groups 3, 4, 5, and 7. Groups 7 and 4 contain primarily *Idaeobatus* species, but they are not closely related. Group 4 is highly supported as sister to Group 3 in analyses of exon sequences for all taxa (**Figure 2**) as well as for diploid taxa only (Carter, 2018). Group 7 further splits into two separate groups in the chloroplast analysis (**Figure 5**). One branch is sister to Group 3 while the other is sister to Group 5, indicating strong maternal genetic differences between these two *Idaeobatus* groups. Multiple studies have recognized polyphyly in *Idaeobatus* (Alice and Campbell, 1999; Morden et al., 2003; Yang and Pak, 2006; Wang et al., 2016). High support for divisions between *Idaeobatus* species in this and other studies indicate that this subgenus would benefit from further phylogenetic study and taxonomic reclassification.

Hybrids

The HybPiper assembly pipeline reduced the complexity of polyploid species by choosing the longest sequence per target locus (Johnson et al., 2016). Because there are hundreds of targets, the evolutionary history of each subgenome in a polyploid was represented by a proportion of the loci, thus, the species trees give a broad overview of that mixed signal. Dichotomous trees place hybrid taxa intermediately between progenitors because their genomes have conflicting phylogenetic signal (Seehausen, 2004). However, if parents are distantly related, the hybrid taxon may not appear phylogenetically close. Without the constraint of dichotomous branching, network analyses allowed a more thorough visualization of such conflict within our data and possible hybrids.

Rubus hybrids "Logan," "Boysen," and "Marion" are horticulturally and economically important cultivars in major berry production regions in the Pacific Northwest and around the world (Jennings, 1988; Thompson, 1997; Hall and Funt, 2017). All three are known blackberry × raspberry hybrids. "Logan" has the closest raspberry relative with 'Red Antwerp' as the documented pollen parent (Jennings, 1988). "Boysen" is an offspring of "Logan" and thus has a raspberry grandparent. "Logan" and "Boysen" are both derived from "Aughinbaugh," a domesticated western North American *R. ursinus* selection (Jennings, 1988). "Marion" has a raspberry for a great-great-grandparent and is also related to *R. ursinus* (Jennings, 1988; Thompson, 1995). All three cultivars cluster with the *R. ursinus* selections in the chloroplast phylogeny, confirming the documented relationships with this species (**Figure 5**). In nuclear analyses, "Logan" groups with other raspberries in Group 7 while "Boysen" and "Marion" are positioned in Group 8 with the blackberries (**Figure 2**). The position of "Logan" with the raspberries is as expected given its paternal red raspberry parent and the possibility that *R. ursinus* may also be a hybrid berry (Alice and Campbell, 1999; Morden et al., 2003). QC values are low or negative for "Boysen" and "Marion" related nodes, indicating that a weak majority or minority of quartets support the position of these species (Carter, 2018). The raspberry germplasm in their recent heritage creates conflict in the phylogenetic signal for these taxa. In network analyses, "Marion" and "Boysen" group with other blackberries in Group 8 while "Logan" is placed within Group 6, between Groups 7 and 8 (**Figure 4**). The placement of "Logan" between Groups 7 (raspberries) and 8 (blackberries) reflects its hybrid heritage.

Evidence of hybridization exists across the *Rubus* phylogeny (**Figure 4A**, **Supplemental Tables S7, S8**). The position of *R. chamaemorus* ($2n = 8x = 56$) (Thompson, 1997) in Group 1 has low support in the exon-based ASTRAL-II phylogeny (**Figure 2**). In a previous study, two *R. chamaemorus* alleles from GBSSI-1γ appeared either outside of the major *Rubus* clade as sister to *R. lasiococcus* or inside as sister to *R. arcticus* (Michael, 2006). *Rubus chamaemorus* may have progenitors outside of and within the major *Rubus* clade, leading to its variable position. The maternal progenitor for *R. chamaemorus* is likely a lineage outside of the major *Rubus* clade since this species is sister to *R. pedatus* in Group 1 in the chloroplast phylogeny (**Figure 5**). This finding supports that *R. chamaemorus* may have an autopolyploid origin (Martinussen et al., 2013).

Rubus humulifolius is strongly associated with Group 4 in the exon ASTRAL-II phylogeny, but groups (with low support) in Group 3 in the chloroplast tree (**Figures 2** and **5**). In the exon split network, *R. humulifolius* occupies a short node between Groups 3 and 4 (**Figure 3**). This indicates that splits in gene trees do not consistently place this species with either Group 3 or Group 4. *Rubus humulifolius* ($2n = 4x = 28$) is the only polyploid taxon in either of these two groups, a trait also indicative of hybrid origin. Progenitors are likely from subg. *Idaeobatus* and/or *Cylactis*.

Similar to *R. humulifolius, R. saxatilis* ($2n = 4x = 28$) is another polyploid in a primarily diploid clade. *Rubus saxatilis* is closely related to subg. *Idaeobatus* species in Group 7, although it is currently classified in subg. *Cylactis*. In the chloroplast tree, this species is sister to the black raspberries, *R. occidentalis, R. leucodermis, R. eriocarpus,* and *R. pungens* (**Figure 5**). Network analyses from exon sequences place *R. saxatilis* between Groups 5 and 7 with a short branch, exhibiting conflict in the placement of this species (**Figure 3**). The supercontigs sequence network places *R. saxatilis* unexpectedly near Group 3 along with *R. caesius* (Carter, 2018). The maternal progenitor of this species is likely from subg. *Idaeobatus*. The paternal parent is unknown and may be a member of Group 3, 5, 6, or 7.

Group 5 members include the Asian polyploids subg. *Malachobatus, Dalibardastrum, Chamaebatus, Cylactis,* and *Idaeobatus*. The diploid exon ASTRAL-II tree shows that Groups 3 and 4 are more closely related to Group 8 than to Group 7 (Carter, 2018). Members of subg. *Idaeobatus*, such as *R. parvifolius* or *R. pentagonus,* and members of subg. *Dalibarda*, such as *R. fockeanus,* may have been progenitors of this likely

allopolyploid subgenus (Wang et al., 2016). *Rubus pentagonus* is closely related to subg. *Malachobatus* species in Group 5, along with other subg. *Idaeobatus* taxa, *R. thomsonii* and the unclassified *R. sengorensis*. The shift in the relationship between Groups 3, 4, 7, and 8 after the addition of putative allopolyploids in Group 5 lends support to the hypothesis that subg. *Malachobatus* is derived from subg. *Idaeobatus* and *Cylactis* species (Wang et al., 2016). Phylogenetic signal from Group 5 brought the progenitor species and their relatives from Groups 3, 4 and 7 together in the dichotomous phylogeny. *Rubus pentagonus, R. thomsonii,* and *R. sengorensis* may be progenitors of this group or examples of subg. *Idaeobatus* hybrids. In the chloroplast analysis, these three species are embedded within Group 5 with other subg. *Malachobatus* and *Dalibardastrum* species. Sister to Group 5 is another group of subg. *Idaeobatus* species, *R. pungens, R. saxatilis, R. eriocarpus, R. occidentalis*, and *R. leucodermis* that could be possible progenitor species.

Species from subg. *Dalibardastrum,* another polyphyletic subgenus in Group 5, are also putative allopolyploids with progenitor species either from or similar to those for subg. *Malachobatus*. Network analyses distinctly show Group 5 separating from other groups, but the extensive webbing between taxa illustrates conflict in the dataset for these species. This demonstrates the convoluted evolutionary history between these putative allopolyploids. Group 5 is positioned between Groups 7 and Groups 3 and 4, which include the proposed progenitors from subg. *Idaeobatus* and *Cylactis* (**Figure 3**).

Blackberry × raspberry hybrids in Group 6 are primarily classified in subg. *Rubus* but are genetically distinct from other blackberries in Group 8 in nuclear analyses. A hybrid subgenus, such as *Idaeorubus* Holub, initially described for cultivars, may be applicable for these taxa.

There are two strongly supported subgroups in Group 8. Subg. *Orobatus* species form one, while Australasian species in subg. *Diemenicus* and *Micranthobatus*, along with southern South American *R. geoides* from subg. *Comaropsis,* form another (**Figure 2**). Both subgroups are distinct from, but closely related to, the major subg. *Rubus* clade. This could be interpreted in two ways. First, populations of the common ancestor of these species may have become reproductively isolated and subsequently evolved into each of these three major groups. It is difficult to reconcile the varying ploidy levels of all species involved with this scenario. Another hypothesis is that both subgroups have one progenitor within or closely related to subg. *Rubus* and another in a different subgenus, such as *Cylactis* for *Comaropsis*/*Diemenicus*/*Micranthobatus* (Jennings, 1988; Hummer and Alice, 2017). The maternal parent in either of these hypothesized crosses is from subg. *Rubus* because all three are in Group 8 in the chloroplast phylogeny (**Figure 5**).

Group 6 contains additional putative hybrids between subg. *Idaeobatus* and subg. *Rubus*. In nuclear phylogenies, this clade shifts positions but is either associated with Group 7 or 8 (**Figure 2**). In the chloroplast phylogeny, these species do not form a clade but all group with subg. *Rubus* in Group 8 (**Figure 5**). The exon network for all taxa places Group 6 between Groups 7 and 8 (**Figure 3**).

Rubus glaucus is morphologically similar to black raspberries (Group 7) with semi-erect, glaucus canes and trifoliate leaves, but has fruit that adheres to the torus like a blackberry (Focke, 1910; Focke, 1911; Focke, 1914; Standley and Steyermark, 1946; Jennings, 1988). It is closely related to black raspberries *R. eriocarpus, R. occidentalis* and *R. leucodermis* in the exon ASTRAL-II phylogeny of all taxa (**Figure 2**). In the chloroplast tree, *R. glaucus* shifts into Group 8 where it is related to subg. *Rubus* and *Orobatus* taxa (**Figure 4**). If it is a cross between a black raspberry and a blackberry, as its morphology suggests and is supported by its variable placement with weak support in nuclear phylogenetic analyses, a black raspberry was likely the paternal donor (Focke, 1910; Focke, 1911; Focke, 1914; Jennings, 1988).

Rubus caesius is a tetraploid blackberry that hybridizes readily with other bramble species (Jennings, 1988; Alice et al., 2001) and has given rise to many new blackberry varieties in Europe (Sochor et al., 2015). The maternal parent for *R. caesius* was likely in subg. *Rubus* given its position in Group 8 in the chloroplast phylogeny (**Figure 4**).

Rubus macraei and *R. hawaiensis* are both endemic Hawaiian species, but are evolutionarily separate. *Rubus hawaiensis* is in Group 3 and sister to *R. spectabilis* with strong support in all analyses (**Figures 2** and **3**). *Rubus macraei,* a hexaploid ($2n = 6x = 42$) (Morden et al., 2003), is a member of Group 6 and another putative blackberry × raspberry hybrid. These results support the hypothesis that *R. hawaiensis* and *R. macraei* arose from separate colonization events of the Hawaiian Islands (Howarth et al., 1997; Morden et al., 2003).

Rubus ursinus is represented by six accessions. Specimens 1 and 6 are putative *R. ursinus* × *armeniacus* hybrids and are in Group 8 in all nuclear analyses (**Figure 2**). In the chloroplast phylogeny, they group with the other *R. ursinus* accessions, indicating that *R. armeniacus* was the pollen parent (**Figure 5**). Despite varying ploidy levels, *R. ursinus* accessions 2, 3, 4, and 5 in Group 6 form a clade (**Figure 2**). Variability in the placement of *R. ursinus* in nuclear phylogenies indicates that the species is a blackberry × raspberry hybrid with the maternal parent in subg. *Rubus* (**Figure 2**) (Carter, 2018). This supports the hypothesis in Alice and Campbell (1999) that *R. ursinus* is a hybrid, however there is no direct evidence that *R. macraei* is a parent of *R. ursinus*. Rather, both of these species are putative blackberry × raspberry hybrids of unknown origin.

Ancestral Ranges and Geographic Migrations

The *Rubus* MRCA is most likely from North America, supporting the hypothesis presented by Alice and Campbell (1999) based on an ITS phylogeny (**Figure 4**). This contradicts hypotheses by Lu (1983) and Kalkman (1988) that *Rubus* originated in southwestern China or Gondwanaland. For *Rubus*, high diversity seen in Asian regions does not correspond with the most likely ancestral range. *Rubus* in

Groups 4, 5, 6 and 7, and 8 colonized Asia at least three times during the Miocene (**Figure 4**). Group 5 is likely the result of a hybridization event between progenitors already distributed in Asia since these species are not present in North America. Groups 4, 7 (both primarily subg. *Idaeobatus*) and 8a (primarily subg. *Rubus*), show classic eastern Asian–eastern North American biogeographic disjunction patterns where closely related species are dispersed across both geographic locations (Nie et al., 2012; Graham, 2018). During the Miocene, plant dispersal from North America to Asia could have occurred over the Bering or North American land bridges (Tiffney, 1985; Wen, 1999; Wen, 2001; Milne, 2006). Distributions in Groups 4 and 7 likely occurred over the Bering land bridge because North American species in these groups are presently distributed in western regions. *Rubus sachalinensis*, an Asian red raspberry, is native to Europe and Asia, but clusters with other North American subg. *Idaeobatus* species, including *R. strigosus*, and the European *R. idaeus*. These European species have a unique evolutionary path compared to other Asian subg. *Idaeobatus* taxa and may be another example of an independent *Idaeobatus* migration from North America into Eurasia. This supports results from Wang (2011) using *matK* chloroplast sequences to study *Rubus* species used in traditional Chinese medicine where *R. sachalinensis* was sister to *Idaeobatus* accessions from Asia. Morphological stasis may explain why character states do not differentiate these genetically differentiated *Idaeobatus* groups. Stasis occurs when evolutionary constraints and stabilizing selection prevent significant changes in morphological characters between lineages (Williamson, 1987; Wen, 2001). This can occur when disjunct geographic areas have similar habitats, such as in North America and eastern Asia (Parks and Wendel, 1990).

In Group 8, the Eurasian distribution of many species and the presence of close genetic relatives in eastern North America suggest migration across the North American land bridge, however this passage closed at the latest 15 Ma (Milne and Abbott, 2002). North American ancestors of Group 8 taxa may have been widespread across North America in the broadleaved, deciduous, temperate forests characterizing the Miocene (Graham, 1993). These species could have migrated across the Bering land bridge through Asia and into Europe. During the subsequent Pleistocene glaciation events, North American distributions shrank back into the east. After diploid species migrated to Europe through the late Miocene, glacial cycles created conditions beneficial for the success of apomictic polyploids. With populations fragmented among glacial refugia, the ability to reproduce asexually may have been advantageous (Sochor et al., 2015).

Ancestors of species distributed in Mexico, Guatemala, and South America, in Groups 2 and 7 (*R. trilobus, R. glaucus,* and *R. eriocarpus*) may have diversified from their North American relatives. This would have occurred after temperature decreases and the spread of grasslands during the Pliocene created refugia of the widespread broadleaved, deciduous forests of the Miocene in the southeastern US and Mexico (Graham, 1993). In the mid-Miocene, the South American subgenus *Orobatus* diverged from other Group 8 taxa. During the Paleogene, approximately 30 Ma, the Panamanian Isthmus connecting Central and South America began to close. The isthmus was crossable for plants and animals at approximately 20 Ma until 3 Ma (O'Dea et al., 2016). *Rubus geoides* in Group 8c also differentiated from North American ancestors during this time frame. Long distance dispersal most clearly explains the disjunction between *R. geoides* in South America and subg. *Micranthobaus/Diemenicus* species in Australia and New Zealand. This vicariance occurs too late (approx. 10 Ma) to have occurred over the land bridge between South America, Antarctica, and Australia, which broke up in the late Eocene approximately 30 Ma, when the continental shelves were no longer exposed (Lawver and Gahagan, 2003). A similar dispersal event occurred in Vitaceae and was likely driven by birds (Nie et al., 2012). Further geographic isolation after dispersal between Tasmania and New Zealand likely led to speciation between *R. parvus* and *R. australis* (New Zealand) and *R. gunnianus* and *R. moorei* (Tasmania) (Hummer and Alice, 2017).

CONCLUSION

Rubus phylogenetic estimation has been complicated by whole genome duplication and hybridization, and informative single-copy nuclear genes have been lacking. Advances in high throughput sequencing now permit hundreds to thousands of loci to be including in a phylogenetic analysis (Weitemier et al., 2014). Our target capture dataset of approximately 1,000 single copy loci provided high resolution between species for many clades but also evidence of gene tree/species tree and cytonuclear discordance. In most cases, discordance is due to biological processes such as hybridization and incomplete lineage sorting as opposed to a lack of phylogenetic signal (Carter, 2018). This study illustrates the importance of using multiple phylogenetic methods when examining complex groups and the utility of software programs that estimate signal conflict within datasets.

The automated analyses, such as HybPiper RogueNaRok, were chosen because they were reliable and repeatable considering the large number of genes and taxa evaluated. Future work could certainly enhance the phylogenetic results through complete taxonomic sampling, longer sequences (PacBio or Nanopore), and by comparing the results to an approach that removes outlier sequences at the alignment stage (Borowiec, 2019). However, these additional analyses are clearly beyond the scope of the current manuscript.

Within each clade, taxon composition and relationships were highly consistent. Differences between datasets and analyses were more evident in the topology of internal nodes delineating the relationships between groups where phylogenetic signal may be obscured by recent polyploidization and hybridization events.

Anoplobatus and *Orobatus* are monophyletic subgenera. Putative allopolyploid subgenera *Dalibardastrum* and *Malachobatus* are closely related and may have progenitors in subg. *Idaeobatus* or *Cylactis*. Subgenus *Idaeobatus* is strongly polyphyletic in nuclear and chloroplast analyses. Subgenus

Rubus is monophyletic with the exception of putative allopolyploids *R. glaucus, R. caesius, and R. ursinus.*

The analysis of cultivated blackberry × raspberry hybrids with known pedigrees confirms the effectiveness of target capture sequencing for phylogenetic analysis. This approach successfully detects and associates hybrid genomes to the appropriate groups. Additional putative hybrids include *R. humulifolius*, with possible parentage from species in subg. *Idaeobatus* and *Cylactis*, and *R. macraei*, with putative progenitors from *Idaeobatus* and a species, such as *R. ursinus*, from subg. *Rubus* (Morden et al., 2003). Long read sequence data and the assembly of haplotypes would give additional insight into difficult-to-classify polyploid, hybrid species like *R. macraei* and *R. chamaemorus* (Kamneva et al., 2017; Dauphin et al., 2018). Haplotype sequencing could allow direct analysis of the evolutionary history of different subgenomes in these putative hybrid species with each subgenome treated as a separate branch on the phylogeny. Instead of hybrids showing an intermediate relationship with progenitors, as in our analysis, subgenome sequences would group directly with parental species. However, our use of hundreds of loci, multiple analysis methods, and assessment of phylogenetic signal supporting internal nodes enabled a critical assessment of the broad evolutionary history of *Rubus.*

Our molecular analysis and dating approach estimated the biogeographical patterns in *Rubus.* The most recent common ancestor was likely distributed in North America. During the early Miocene, lineages likely migrated from North America to Asia and Europe over the Bering land bridge. Migrations to South America occurred during the formation of the Panamanian Isthmus in the mid- to late Miocene, and long-distance dispersal events may have allowed *Rubus* to spread from South America to Australia and New Zealand. During the middle and late Miocene the genus diversified greatly in Asia, Europe, South America and Oceania. Whole genome duplication events occurred producing higher ploidy species on multiple continents. Cooling temperatures and glaciation isolated Central American populations from North America, and may have created conditions beneficial to the formation of apomictic polyploids in Europe. While our research sets the stage for reassessing *Rubus* subtaxa, i.e., subgenera or sections, a thorough morphological evaluation of multiple accessions of species across the genus must follow to identify useful synapomorphies for taxonomic redefinition.

DATA AVAILABILITY STATEMENT

The datasets generated for this study, such as sequence alignments and phylogenies, are available at the OSU scholars archive https://ir.library.oregonstate.edu/concern/file_sets/6108vh40r. Reads are available in the NCBI Short Read Archive (SRA): PRJNA510412.

AUTHOR CONTRIBUTIONS

KC contributed to the laboratory work, data analyses, and manuscript writing. LA, BS, and JB contributed to the laboratory work and manuscript review. TM and DB contributed to data analysis and manuscript review. AL, LA, NB, and KH conceived the study and contributed to the analysis and manuscript preparation. All authors have read and approved the final manuscript.

FUNDING

This work was supported by USDA ARS CRIS 2072-21000-044-00D and 2072-21000-049-00D and NSF KY EPSCoR National Laboratory Initiative 019-14 and NSF DEB award to LA for this research.

ACKNOWLEDGMENTS

We appreciate the technical assistance of the Center for Genome Research and Biocomputing at Oregon State University for Illumina® sequencing. We thank M. Dossett, R. Cronn, K. Weitemier, R. Schmickl, J.C. Lee, R. Meiers, C. Mulch, A.M. Nyberg, M. Peterson, M. Clark, K.J. Vining, M.L. Worthington, M.H. Yin, J.D. Zurn, J.R. Clark, and C.E. Finn for technical assistance and meaningful discussion on this manuscript. This manuscript has been released as a pre-print at biorXiv (Carter et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01615/full#supplementary-material

REFERENCES

Aberer, A. J., Krompass, D., and Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst. Biol.* 62, 162–166. doi: 10.1093/sysbio/sys078

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle,[w]: *Proceedings of the 2nd International Symposium on Information*. Eds. B. N. Petrow and F. Czaki (Budapest: Akademiai Kiado).

Alice, L. A., and Campbell, C. S. (1999). Phylogeny of Rubus (Rosaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. *Am. J. Bot.* 86, 81–97. doi: 10.2307/2656957

Alice, L. A., Eriksson, T., Eriksen, B., and Campbell, C. S. (2001). Hybridization and gene flow between distantly related species of Rubus (Rosaceae): evidence from nuclear ribosomal DNA internal transcribed spacer region sequences. *Syst. Bot.* 26, 769–778. doi: 10.1043/0363-6445-26.4.769

Alice, L., Dodson, T., and Sutherland, B. (2008). Diversity and relationships of Bhutanese Rubus. *Acta. Horticulturae*. 777, 63–70. doi: 10.17660/ActaHortic.2008.777.5

Alice, L. A., Goldman, D. H., Macklin, J. A., and Moore, G. (2015). Rubus. *In Flora of North America Editorial Committee, The Flora of North America (Rosaceae)* (Missouri Botanical Garden Press), 9, 28–56.

Andersen, M. J., McCullough, J. M., Mauck, W. M. III, Smith, B. T., and Moyle, R. G. (2018). A phylogeny of kingfishers reveals an Indomalayan origin and elevated rates of diversification on oceanic islands. *J. Biogeogr.* 45 (2), 269–281. doi: 10.1111/jbi.13139

Bailey, L. H. (1941). *Species batorum: the genus Rubus in North America* (Bailey Hortorium of the The New York State College of Agriculture at Cornell University).

Bammi, R. K., and Olmo, H. P. (1966). Cytogenetics of Rubus. v. natural hybridization between R. procerus PJ Muell. and R. laciniatus willd. *Evolution* 20, 617–633. doi: 10.1111/j.1558-5646.1966.tb03392.x

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.20120021

Barneby, R. C. (1988). Flora of bhutan, including a record of plants from Sikkim. *Brittonia* 1, 289–289. doi: 10.2307/2807475 Part 340.

Bean, A. R. (1995). Revision of *Rubus* subgenus *Micranthobatus* (Fritsch) Kalkman (Rosaceae) in Australia. *Austrobaileya* 4, 321–328.

Bean, A. R. (1997). A revision of Rubus subg. Malachobatus (Focke) Focke and Rubus subg. Diemenicus AR Bean (Rosaceae) in Australia. *Austrobaileya*, 39–51.

Borowiec, M. L. (2019). Spruceup: fast and flexible identification, visualization, and removal of outliers from large multiple sequence alignments. *J. Open Source Software* 4 (42), 1635. doi: 10.21105/joss.01635

Bushnell, B. (2014). BBTools software package. URL http://sourceforgenet/projects/bbmap.

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

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. *bioRxiv*, 703926. doi: 10.1101/703926

Carter, K. A. (2018), *Phylogenetic estimation and ancestral state reconstruction of Rubus (Rosaceae) using target capture sequencing In Department of Horticulture* (Corvallis, Oregon: Oregon State University), 238.

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

Dauphin, B., Grant, J. R., Farrar, D. R., and Rothfels, C. J. (2018). Rapid allopolyploid radiation of moonwort ferns (Botrychium; Ophioglossaceae) revealed by PacBio sequencing of homologous and homeologous nuclear regions. *Molecular Phylogenet. Evol.* 120, 342–353. doi: 10.1016/j.ympev.2017.11.025

Dillenberger, M. S., Wei, N., Tennessen, J. A., Ashman, T. L., and Liston, A. (2018). Plastid genomes reveal recurrent formation of allopolyploid Fragaria. *Am. J. Bot.* 105 (5), 862–874. doi: 10.1002/ajb21085

Doyle, J., and Doyle, J. L. (1987). Genomic plant DNA preparation from fresh tissue-CTAB method. *Phytochem. Bull.* 19, 11–15.

FAO. (2019). Food and Agriculture Organization of the United Nations. FAOSTAT. Available at: http://www.fao.org/faostat/en/#data/QC accessed 12/03/2019

Focke, W. O. (1910). Species ruborum. *Monographiae generis rubi prodromus* Vol. 17 (Stuttgart: Stuttgart, E. Schweizerbart).

Focke, W. O. (1911). Species ruborum. *Monographiae generis rubi prodromus* Vol. 17 (Stuttgart: Stuttgart, E. Schweizerbart).

Focke, W. O. (1914). Species ruborum. *Monographiae generis rubi prodromus* Vol. 17 (Stuttgart: Stuttgart, E. Schweizerbart).

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

Gleason, H., and Cronquist, A. (1991). *Manual of vascular plants of northeastern North America and adjacent Canada* (Bronx, New York, USA: New York Botanical Garden).

Graham, A. (1993). History of the vegetation: Cretaceous (Maastrichtian)-Tertiary. *Flora North Am.* 1, 57–70.

Graham, A. (2018). The role of land bridges, ancient environments, and migrations in the assembly of the North American flora. *J. Syst. Evol.* 56 (5), 405–429. doi: 10.1111/jse.12302

Grunewald, S., Spillner, A., Bastkowski, S., Bogershausen, A., and Moulton, V. (2013). SuperQ: computing supernetworks from quartets. *IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB)* 10, 151–160. doi: 10.1109/TCBB.2013.8

Hall, H. K., and Funt, R. C. (2017). Blackberries and their hybrids. *Crop Prod. Sci. Hortic.* (Oxfordshire: United, Kingdom) 26. doi: 10.1079/97817806466880000 (CABI).

Holmgren, P. K., Holmgren, N. H., and Barnett, L. C. (1990). *Index Herbariorum, Part I: The herbaria of the world* (New York, New York: New York Botanical Garden).

Howarth, D. G., Gardner, D. E., and Morden, C. W. (1997). Phylogeny of Rubus subgenus Idaeobatus (Rosaceae) and its implications toward colonization of the Hawaiian Islands. *Syst. Bot.* 22, 433–441. doi: 10.2307/2419819

Hummer, K. E., and Alice, L. A. (2017). Small genomes in tetraploid Rubus L. (Rosaceae) from New Zealand and Southern South America. *J. Am. Pomol. Soc.* 71, 2–7.

Hummer, K. E., Bassil, N. V., and Alice, L. A. (2015). Rubus ploidy assessment. *Acta Hortic.* 1133, 81–88. doi: 10.17660/ActaHortic.2016.1133.13

Hytönen, T., Graham, J., and Harrison, R. (2018). *The Genomes of Rosaceous Berries and Their Wild Relatives* (Switzerland AG: Springer Nature). doi: 10.1007/978-3-319-76020-9

ITPGR. (2019). *International Treaty on Plat Genetic Resources for Food and Agriculture* (Food and Agriculture Organization of the United Nations).

Jennings, D. L. (1988). *Raspberries and blackberries: their breeding, diseases and growth* (London, UK: Academic Press).

Johnson, M. G., Gardner, E. M., Yang, L., 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, 1–7. doi: 10.3732/apps.1600016 apps.1600016.

Kalkman, C. (1988). The phylogeny of the Rosaceae. *Bot. J. Linn. Soc.* 98, 37–59. doi: 10.1111/j.1095-8339.1988.tb01693.x

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

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

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

Lawver, L. A., and Gahagan, L. M. (2003). Evolution of Cenozoic seaways in the circum-Antarctic region. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 198, 11–37. doi: 10.1016/S0031-0182(03)00392-4

Leavitt, S. D., Kirika, P. M., Depaz, G. A., Huang, J. P., Jae-Seoun, H. U. R., Grewe, F., et al. (2018). Assessing phylogeny and historical biogeography of the largest genus of lichen-forming fungi, Xanthoparmelia (Parmeliaceae, Ascomycota). *Lichenol.* 50 (3), 299–312. doi: 10.1017/S0024282918000233

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

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint. *arXiv*, 13033997.

Liston, A. (2014). 257 nuclear genes for Rosaceae phylogenomics.

Lu, L., and Boufford, D. E. (2003). "Rubus L," in *Flora China*. 9, 195–286.

Lu, Y. Y., He, K., Klaus, S., Brown, R. M., and Li, J. T. (2018). A comprehensive phylogeny of the genus Kurixalus (Rhacophoridae, Anura) sheds light on the geographical range evolution of frilled swamp treefrogs. *Mol. Phylogenet. Evol.* 121, 224–232. doi: 10.1016/j.ympev.2017.09.019

Lu, L.-T. (1983). Study on the genus Rubus of China. *Chih wu fen lei hsueh pao = Acta. Phytotaxonomica Sin.* 21 (1), 13–25.

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.* 46, 523–536. doi: 10.1093/sysbio/46.3.523

Martinussen, I., Uleberg, E., Sønsteby, A., Sønstebø, J. H., Graham, J., and Vivian-Smith, A. (2013). Genomic survey sequences and the structure of the Rubus chamaemorus L. genome as determined by ddRAD tags.

Matzke, N. J. (2013) BioGeoBEARS: biogeography with Bayesian (and likelihood) evolutionary analysis in R scripts. R package, version. 11 1, 2013.

Matzke, N. J. (2014). Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. *Syst. Biol.* 63, 951–970. doi: 10.1093/sysbio/syu056

Meng, R., and Finn, C. E. (2002). Determining ploidy level and nuclear DNA content in Rubus by flowcytometry. *J. Am. Soc. Hortic. Sci.* 127, 767–775. doi: 10.21273/JASHS.127.5.767

Michael, K. (2006). "Clarification of basal relationships in Rubus (Rosaceae) and the origin of Rubus chamaemorus," in *Department of Biology* (Bowling Green, Kentucky: Westren Kentucky University).

Milne, R. I., and Abbott, R. J. (2002). The origin and evolution of Tertiary relict floras. *Adv. Bot. Res.* 38, 281–314. doi: 10.1016/S0065-2296(02)38033-9

Milne, R. I. (2006). Northern hemisphere plant disjunctions: a window on Tertiary land bridges and climate change? *Ann. Bot.* 98, 465–472. doi: 10.1093/aob/mcl148

Mimura, M., Mishima, M., Lascoux, M., and Yahara, T. (2014). Range shift and introgression of the rear and leading populations in two ecologically distinct Rubus species. *BMC Evol. Biol.* 14, 209. doi: 10.1186/s12862-014-0209-9

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

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

Morden, C. W., Caraway, V., and Motley, T. J. (1996). Development of a DNA library for native Hawaiian plants. *Pac. Sci.* 50, 324–335.

Morden, C. W., Gardner, D. E., and Weniger, D. A. (2003). Phylogeny and biogeography of pacific Rubus subgenus Idaeobatus (Rosaceae) species: Investigating the origin of the endemic Hawaiian raspberry R. macraei. *Pac. Sci.* 57, 181–197. doi: 10.1353/psc.20030018

Nie, Z.-L., Sun, H., Manchester, S. R., Meng, Y., Luke, Q., and Wen, J. (2012). Evolution of the intercontinental disjunctions in six continents in the Ampelopsis clade of the grape family (Vitaceae). *BMC Evol. Biol.* 12, 17–17. doi: 10.1186/1471-2148-12-17

O'Dea, A., Lessios, H. A., Coates, A. G., Eytan, R. I., Restrepo-Moreno, S. A., Cione, A. L., et al. (2016). Formation of the Isthmus of Panama. *Sci. Adv.* 2, e1600883. doi: 10.1126/sciadv.1600883

Pankhurst, R. (2001). Rosaceae. In: W.D. Stevens, C. Ulloa, A. Pool and O. M. Montiel (eds.). *Flora de Nicaragua. Monographs in Systematic Botany from the Missouri Botanical Garden* 81, 2202–2206.

Parks, C. R., and Wendel, J. F. (1990). Molecular divergence between Asian and North American species of Liriodendron (Magnoliaceae) with implications for interpretation of fossil floras. *Am. J. Bot.* 77, 1243–1256. doi: 10.1002/j.1537-2197.1990.tb11376.x

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/ajb21016

Potter, D., Eriksson, T., Evans, R., Oh, S.-H., Smedmark, J. E. E., Morgan, R. D., et al. (2007). Phylogeny and classification of Rosaceae. *Plant Sys. Evol.* 266, 5–43. doi: 10.1007/s00606-007-0539-9

Quinlan, A. R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinf.* 47, 11.12. doi: 11-11.12.34.10.1002/0471250953.bi1112s47

Reaz, R., Bayzid, M. S., and Rahman, M. S. (2014). Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLoS One* 9, e104008. doi: 10.1371/journal.pone.0104008

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

Romoleroux, K., Nllgaard, B., Harling, G., and Andersoon, L. (1996). Flora of Ecuador: 79. Rosaceae; 81. Connaraceae (Department of Systematic Botany).

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* 19, 101–109. doi: 10.1093/oxfordjournals.molbev.a003974

Sanderson, M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302. doi: 10.1093/bioinformatics/19.2.301

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

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends. Ecol. Evol.* 19, 198–207. doi: 10.1016/j.tree.2004.01.003

Sochor, M., Vašut, R. J., Sharbel, T. F., and Trávníček, B. (2015). How just a few makes a lot: speciation *via* reticulation and apomixis on example of European brambles (Rubus subgen. Rubus, Rosaceae). *Mol. Phylogenet. Evol.* 89, 13–27. doi: 10.1016/j.ympev.2015.04.007

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

Standley, P. C., and Steyermark, J. A. (1946). *Flora of Guatemala* (Flora of Guatemala). Natural History Museum, IL. 503.

Sutherland, B. (2005). Phylogenetics of Rubus ursinus and R. macraai (Rosaceae): Evidence of hybrid origin. *Stud. Honors Theses* 1–186.

Swofford, D. (2003). PAUP* ver 4.0. b10. Phylogenetic Analysis Using Parsimony and Other Methods Sunderland, MA: Sinauer Associates, Sunderland.

Thompson, M. M. (1995). Chromosome numbers of Rubus species at the national clonal germplasm repository. *Hort. Sci.* 30, 1447–1452. doi: 10.21273/HORTSCI.30.71447

Thompson, M. M. (1997). Survey of chromosome numbers in Rubus (Rosaceae: Rosoideae). *Ann. Missouri Bot. Garden*, 128–164. doi: 10.2307/2399958

Tiffney, B. H. (1985). Perspectives on the origin of the floristic similarity between eastern Asia and eastern North America. *J. Arnold Arboretum* 66, 73–94. doi: 10.5962/bhl.part.13179

VanBuren, R., Bryant, D., Bushakra, J. M., Vining, K. J., Edger, P. P., Rowley, E. R., et al. (2016). The genome of black raspberry (Rubus occidentalis). *Plant J.* 87, 535–547. doi: 10.1111/tpj.13215

Veitia, R. A. (2005). Paralogs in polyploids: one for all and all for one? *Plant Cell.* 17 (1), 4–11 doi: 10.1105/tpc.104.170130

Wang, Y., Wang, X., Chen, Q., Zhang, L., Tang, H., Luo, Y., et al. (2015). Phylogenetic insight into subgenera Idaeobatus and Malachobatus (Rubus, Rosaceae) inferring from ISH analysis. *Mol. Cytogenet.* 8, 11. doi: 10.1186/s13039-015-0114-y

Wang, Y., Chen, Q., Chen, T., Tang, H., Liu, L., and Wang, X. (2016). Phylogenetic insights into Chinese Rubus (Rosaceae) from multiple chloroplast and nuclear DNAs. *Front. Plant Sci.* 7, 1–13. doi: 10.3389/fpls.2016.00968

Wang, Y. (2011). Relationships among Rubus (Rosaceae) species used in traditional Chinese medicine. *Masters Theses and Specialists Project*. Paper 1073

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. *Appl. Plant Sci.* 2, 1400042. doi: 10.3732/apps.1400042

Wen, J., Nie, Z.-L., and Ickert-Bond, S. M. (2016). Intercontinental disjunctions between Eastern Asia and Western North America in vascular plants highlight the biogeographic importance of the Bering land bridge from late Cretaceous to Neogene. *J. Syst. Evol.* 54, 469–490. doi: 10.1111/jse.12222

Wen, J. (1999). Evolution of Eastern Asian and Eastern North American disjunct distributions in flowering plants. *Ann. Rev. Ecol. Syst.* 30, 421–455. doi: 10.1146/annurev.ecolsys.30.1.421

Wen, J. (2001). Evolution of Eastern Asian and Eastern North American biogeographic disjunctions: a few additional issues. *Int. J. Plant Sci.* 162, S117–S122. doi: 10.1086/322940

Wilkinson, M., and Crotti, M. (2017). Comments on detecting rogue taxa using RogueNaRok. *Syst. Biodivers.* 15 (4), 291–295.

Williamson, P. G. (1987). *Selection or constraint? A proposal on the mechanism for stasis* (London: Rates of evolution Allen and Unwin), 129–142. doi: 10.4324/9780429293849-6

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T. S., et al. (2016). 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

Yang, J. Y., and Pak, J.-H. (2006). Phylogeny of Korean Rubus (Rosaceae) based on ITS (nrDNA) and trnL/F intergenic region (cpDNA). *J. Plant Biol.* 49, 44–54. doi: 10.1007/BF03030787

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

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: 09 January 2020
doi: 10.3389/fpls.2019.01655

Tackling Rapid Radiations With Targeted Sequencing

Isabel Larridon[1,2†], Tamara Villaverde[3,4,5*†], Alexandre R. Zuntini[1], Lisa Pokorny[1,3,6], Grace E. Brewer[1], Niroshini Epitawalage[1], Isabel Fairlie[1,7], Marlene Hahn[4], Jan Kim[1], Enrique Maguilla[4,8], Olivier Maurin[1], Martin Xanthos[1], Andrew L. Hipp[4,5], Félix Forest[1] and William J. Baker[1]*

[1] Royal Botanic Gardens, Kew, Surrey, United Kingdom, [2] Systematic and Evolutionary Botany Lab, Department of Biology, Ghent University, Ghent, Belgium, [3] Real Jardín Botánico (RJB-CSIC), Madrid, Spain, [4] The Morton Arboretum, Lisle, IL, United States, [5] The Field Museum, Chicago, IL, United States, [6] Centre for Plant Biotechnology and Genomics (CBGP, UPM-INIA), Madrid, Spain, [7] Department of Animal and Plant Sciences, University of Sheffield, Sheffield, United Kingdom, [8] Departamento de Biología Vegetal y Ecología, Universidad de Sevilla, Sevilla, Spain

OPEN ACCESS

Edited by:
Michael R. McKain,
University of Alabama,
United States

Reviewed by:
Angela Jean McDonnell,
Chicago Botanic Garden,
United States
Ryan Folk,
Mississippi State University,
United States

***Correspondence:**
Isabel Larridon
i.larridon@kew.org
Tamara Villaverde
t.villaverde@rjb.csic.es

[†]These authors share first authorship

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

Received: 04 July 2019
Accepted: 22 November 2019
Published: 09 January 2020

Citation:
Larridon I, Villaverde T, Zuntini AR, Pokorny L, Brewer GE, Epitawalage N, Fairlie I, Hahn M, Kim J, Maguilla E, Maurin O, Xanthos M, Hipp AL, Forest F and Baker WJ (2020) Tackling Rapid Radiations With Targeted Sequencing. Front. Plant Sci. 10:1655. doi: 10.3389/fpls.2019.01655

In phylogenetic studies across angiosperms, at various taxonomic levels, polytomies have persisted despite efforts to resolve them by increasing sampling of taxa and loci. The large amount of genomic data now available and statistical tools to analyze them provide unprecedented power for phylogenetic inference. Targeted sequencing has emerged as a strong tool for estimating species trees in the face of rapid radiations, lineage sorting, and introgression. Evolutionary relationships in Cyperaceae have been studied mostly using Sanger sequencing until recently. Despite ample taxon sampling, relationships in many genera remain poorly understood, hampered by diversification rates that outpace mutation rates in the loci used. The C4 *Cyperus* clade of the genus *Cyperus* has been particularly difficult to resolve. Previous studies based on a limited set of markers resolved relationships among *Cyperus* species using the C3 photosynthetic pathway, but not among C4 *Cyperus* clade taxa. We test the ability of two targeted sequencing kits to resolve relationships in the C4 *Cyperus* clade, the universal Angiosperms-353 kit and a Cyperaceae-specific kit. Sequences of the targeted loci were recovered from data generated with both kits and used to investigate overlap in data between kits and relative efficiency of the general and custom approaches. The power to resolve shallow-level relationships was tested using a summary species tree method and a concatenated maximum likelihood approach. High resolution and support are obtained using both approaches, but high levels of missing data disproportionately impact the latter. Targeted sequencing provides new insights into the evolution of morphology in the C4 *Cyperus* clade, demonstrating for example that the former segregate genus *Alinula* is polyphyletic despite its seeming morphological integrity. An unexpected result is that the *Cyperus margaritaceus-Cyperus niveus* complex comprises a clade separate from and sister to the core C4 *Cyperus* clade. Our results demonstrate that data generated with a family-specific kit do not necessarily have more power than those obtained with a universal kit, but that data generated with different targeted sequencing kits can often

be merged for downstream analyses. Moreover, our study contributes to the growing consensus that targeted sequencing data are a powerful tool in resolving rapid radiations.

Keywords: C4 *Cyperus* clade, Cyperaceae, Plant and Fungal Trees of Life, phylogenomics, polytomy, targeted sequencing

INTRODUCTION

Since the late 1980s, molecular phylogenetics has yielded major new insights into the evolution of land plants, especially for flowering plants (e.g., Chase et al., 1993; Ruhfel et al., 2014; Wickett et al., 2014; APG IV, 2016). However, uncertainty in topologies has persisted, particularly for deep nodes (Wickett et al., 2014) and for ancient and recent rapid radiations, which are often inferred as polytomies (Fishbein et al., 2001; Whitfield and Lockhart, 2007; Snak et al., 2016; Spalink et al., 2016). Researchers have attempted to resolve these issues by increasing taxon sampling, the number of DNA loci sampled, or both (e.g., Philippe et al., 2011; Nabhan and Sarkar, 2012; Nicholls et al., 2015).

Targeted sequencing of genomic libraries can yield hundreds to thousands of DNA loci across multiple individuals and species, depending on the targeted sequencing kit used (e.g., Faircloth et al., 2018; Johnson et al., 2018; Couvreur et al., 2019), providing sequencing data suitable to addressing challenging and outstanding problems in plant systematics. It is an extremely versatile technique that can be used to solve ancient and recent species radiations (Nicholls et al., 2015; Stevens et al., 2015; Mitchell et al., 2017; Kadlec et al., 2017), as well as to bridge micro- and macroevolutionary levels (Kates et al., 2018; Villaverde et al., 2018). Additionally, it works well with degraded DNA template, e.g., herbarium material (Hart et al., 2016; McKain et al., 2018; Brewer et al., 2019). Targeted sequencing is rapidly becoming a standard phylogenomic method for flowering plants (McKain et al., 2018).

Evolutionary relationships in the sedge family (Cyperaceae) have mainly been studied using Sanger sequencing (e.g., Simpson et al., 2007; Muasya et al., 2009; Jung and Choi, 2012; Escudero and Hipp, 2013; Spalink et al., 2016; Semmouri et al., 2019; **Figure**

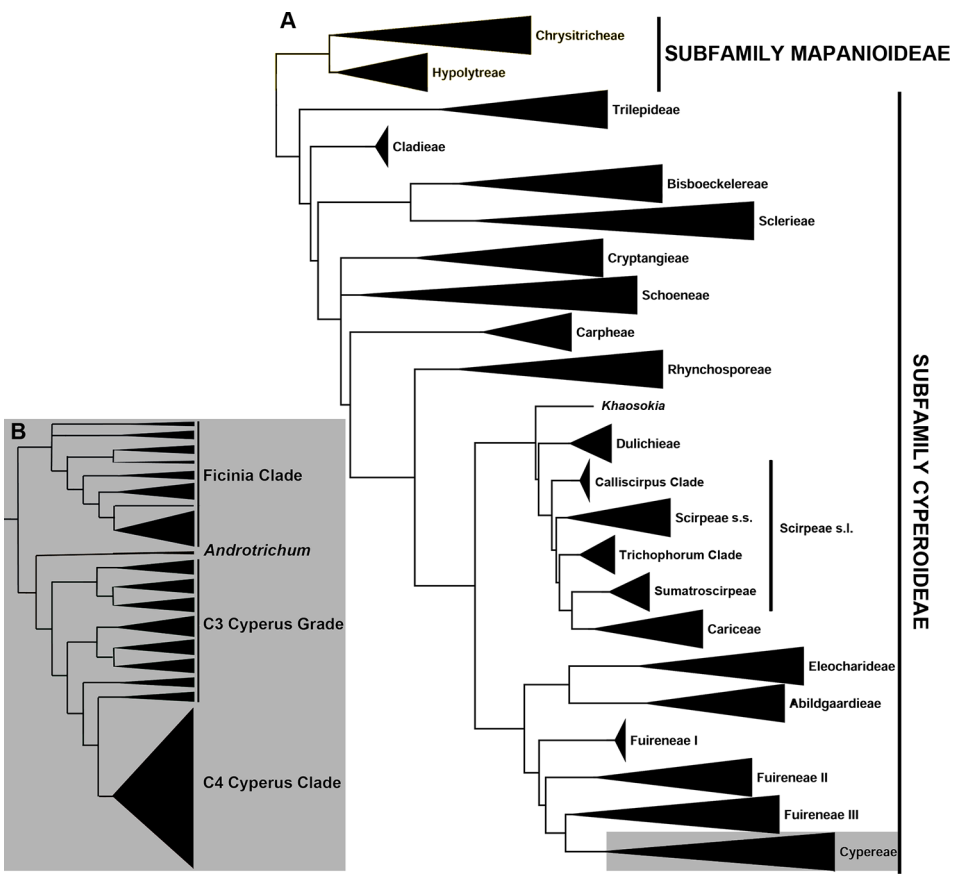

FIGURE 1 | Phylogenetic relationships in Cyperaceae **(A)** Summary phylogeny of the family Cyperaceae based on Semmouri et al. (2019). **(B)** Summary phylogeny of the tribe *Cypereae* based on Larridon et al. (2011a, 2013).

1A). To date, high-throughput sequencing approaches in sedges include a targeted sequencing study using anchored hybrid enrichment or anchored phylogenomics (Lemmon et al., 2012; Buddenhagen et al., 2016), focusing on the Scirpo-Caricoid clade (Léveillé-Bourret et al., 2018), and two other studies using reduced-representation phylogenomic methods, restriction-site associated DNA sequencing (RAD-Seq; Baird et al., 2008), and genotyping-by-sequencing (GBS; Elshire et al., 2011; Deschamps et al., 2012), to resolve fine-scale relationships in the megadiverse genus *Carex* (Escudero et al., 2014; Maguilla et al., 2017). Hyb-Seq, i.e., targeted sequencing combined with genome skimming (Weitemier et al., 2014; Dodsworth et al., 2019), was recently used to investigate the broad-scale relationships in *Carex* (Villaverde et al., in review). Additionally, relationships at tribal and generic levels in Cyperaceae are being investigated using targeted sequencing (I. Larridon et al., unpubl. data). Relationships among Cyperaceae taxa that use the C4 photosynthetic pathway remain ill-understood, hampered by an apparent faster rate of diversification leading to limited topological resolution (e.g., Larridon et al., 2013; Bauters et al., 2014; Roalson et al., 2019). In particular, the relationships between C4 *Cyperus* L. species are still unresolved.

Within tribe *Cypereae*, the *Cyperus* clade includes two genera, i.e., the giant genus *Cyperus* (950 species) and the small genus *Androtrichum* (Brongn.) Brongn. (two species). Thirteen segregate genera recognized by Goetghebeur (1998) have since been subsumed into *Cyperus* (Larridon et al., 2011a; Larridon et al., 2011b; Larridon et al., 2013; Larridon et al., 2014; Bauters et al., 2014), comprising three genera that use C3 photosynthesis —*Courtoisina* Soják, *Kyllingiella* R.W. Haines & Lye, and *Oxycardium* Nees—and 10 genera that use C4 photosynthesis —*Alinula* J. Raynal, *Ascolepis* Nees ex Steud., *Ascopholis* C.E.C. Fisch., *Kyllinga* Rottb., *Lipocarpha* R. Br., *Pycreus* P. Beauv., *Queenslandiella* Domin, *Remirea* Aubl., *Sphaerocyperus* Lye, and *Volkiella* Merxm. & Czech. The small genus *Androtrichum* (C3 photosynthesis) has not yet been subsumed into *Cyperus* because only *rbcL* sequences are available to date, and the phylogenetic placement of the genus is consequently unresolved (Muasya et al., 2009; Jung and Choi, 2012; Hinchliff and Roalson, 2013; Semmouri et al., 2019). Given the ecological importance of *Cyperus*, which achieves high diversity and biomass in ecoregions across the tropics (Larridon et al., 2013; Kipkemboi and van Dam, 2018), and its ethnobotanical significance (Simpson and Inglis., 2001), understanding the diversity of the clade is of high societal importance.

Previous studies based on a limited set of plastid and nuclear ribosomal DNA (nrDNA) markers resolved relationships among species of *Cyperus* using the C3 photosynthetic pathway (Larridon et al., 2011a; Larridon et al., 2011b), but not among sections and species using C4 photosynthesis (Larridon et al., 2013; Larridon et al., 2014; Bauters et al., 2014). In these studies, the species-poor lineages of genus *Cyperus*, which diverge from deeper nodes, form a grade of generally well-circumscribed *Cyperus* sections that all use C3 photosynthesis (hereafter C3 *Cyperus* grade, c. 190 species); while the more highly derived clade (hereafter C4 *Cyperus* clade) represents a radiation of c. 760 *Cyperus* species that use C4 photosynthesis (**Figure 1B**). The C4 *Cyperus* clade is a particularly challenging lineage taxonomically (Huygh et al., 2010; Larridon et al., 2011c; Reynders et al., 2011; Bauters et al., 2014) and previous attempts to resolve relationships within it have resulted in a polytomy (Muasya et al., 2002; Larridon et al., 2011a; Larridon et al., 2013; Bauters et al., 2014). Spalink et al. (2016) showed an increased diversification rate for the C4 *Cyperus* clade commencing c. 20 Ma (million years ago).

It is still unclear under what conditions universal targeted sequencing kits, which target low-copy nuclear markers conserved across a wide phylogenetic range (e.g., angiosperms; Buddenhagen et al., 2016), can be used to infer relationships in fast evolving lineages. If they can, then there may be little benefit to developing custom probes for studies of small numbers of taxa, and there are obvious downstream benefits in using universal probes, data from which can be readily combined across labs. A recent study (Kadlec et al., 2017) on the heather genus *Erica* (Ericaceae, with well over 800 species) concluded that data from markers that are custom-designed using existing pipelines (e.g., MarkerMiner; Chamala et al., 2015) may deliver better results than those obtained using a more universal approach. As Liu et al. (2019) have shown, capture success drops significantly when probe sequences used in a targeted sequencing kit diverge >30% from their intended targets (**Supplementary Figures 5** and **6** in Liu et al., 2019). Recently, an angiosperm-wide targeted sequencing kit, i.e., Angiosperms-353 (Johnson et al., 2018), has been designed using a k-medoids clustering algorithm (Bauckhage, 2015) from a much larger dataset, including several published genomes (available from https://phytozome.jgi.doe.gov) plus 655 angiosperm transcriptomes generated by the one thousand plant transcriptomes (1KP) initiative (Matasci et al., 2014), in the context of the Plant and Fungal Trees of Life (PAFTOL) research program at Royal Botanic Gardens, Kew (https://www.kew.org/science/who-we-are-and-what-we-do/strategic-outputs-2020/plant-and-fungal-trees-life). The Angiosperms-353 kit (Johnson et al., 2018) targets 353 putatively single-copy protein-coding genes (spanning 260,802 bp in total) and was designed using transcriptome data from representatives of all major clades in angiosperms (three accessions belong in Cyperaceae—*Cyperus*, *Lepidosperma*, and *Mapania*—out of 128 monocots in total), to keep expected divergence between all potential taxa and the probes below the 30% divergence threshold beyond which capture is no longer efficient, as Liu et al. (2019) experimentally determined. This kit includes multiple probes for each locus (3x tiling) to optimize its performance with low-quality template (e.g., historical herbarium collections; Brewer et al., 2019). The aforementioned reasons may result in the Angiosperms-353 kit being more successful than other universal targeted sequencing kits for flowering plants.

Here, we present novel data from the C4 *Cyperus* clade obtained using both the Angiosperms-353 targeted sequencing kit and a Cyperaceae-specific kit designed by Villaverde et al. (in review) using the MarkerMiner pipeline (Chamala et al., 2015), with transcriptome data for *Cyperus papyrus* L. (1KP) and *Carex*

siderosticta Hance (S. Kim et al., unpubl. data). The Cyperaceae-specific kit targets 554 low-copy nuclear orthologous loci, spanning c. 1 Mbp. We use these data to: 1) test the effectiveness of the Angiosperms-353 kit to resolve hitherto intractable relationships, 2) compare the relative effectiveness of these universal probes to the Cyperaceae-specific probes, and 3) establish well-supported relationships among this ecologically important group of sedges.

MATERIALS AND METHODS

Taxon Sampling

Sampling for enrichment with the Angiosperms-353 kit consisted of 38 *Cyperus* accessions (one C3 *Cyperus* species, i.e., *Cyperus kyllingiella* Larridon, and 37 species from the C4 *Cyperus* clade) (**Supplementary Table 1A**). Sampling for enrichment with the Cyperaceae-specific kit consisted of eight species of the C4 *Cyperus* clade and *Schoenoplectus pungens* (Vahl) Palla (tribe Fuireneae) used as outgroup (**Supplementary Table 1B**). *Cyperus esculentus* L., *Cyperus mindorensis* (Steud.) Huygh, and *Cyperus richardii* Steud. were enriched with both kits.

Deoxyribonucleic Acid Extraction, Library Preparation, Hybridization, and Sequencing

The voucher information and treatment of each accession is provided (**Supplementary Tables 1** and **2**). Both the Angiosperms-353 (Johnson et al., 2018) and the Cyperaceae-specific (Villaverde et al. in review) kits are available from Arbor Biosciences (Ann Arbor, MI, USA).

Molecular work for accessions enriched with the Angiosperms-353 probes was carried out at the Sackler Phylogenomic Laboratory, within the Jodrell Laboratory at Royal Botanic Gardens, Kew (Richmond, Surrey, UK). Genomic DNA was extracted from leaf tissue obtained from herbarium specimens or silica collected samples, using either a modified cetyl trimethylammonium bromide (CTAB) approach (Doyle and Doyle, 1987) or a CTAB protocol, based on Beck et al. (2012), modified for optimal simultaneous extraction of 96 to 192 samples (i.e., one or two plates) from suboptimal (i.e., herbarium) tissue (Fairlie & Pokorny protocol provided in **Supplementary Data Sheet 1**). Lastly, two accessions were sourced from the Kew DNA Bank (http://dnabank.science.kew.org/) (**Supplementary Table 1A**). The samples extracted using a CTAB approach were purified using Agencourt AMPure XP Bead Clean-up (Beckman Coulter, Indianapolis, IN, USA). All DNA extracts were quantified using a Quantus™ Fluorometer (Promega Corporation, Madison, WI, USA) and then run on a 1% agarose gel to assess the average fragment size. Samples with very low concentration (not visible on a 1% agarose gel), were assessed on an Agilent Technologies 4200 TapeStation System using Genomic DNA ScreenTape (Santa Clara, CA, USA). DNA extracts with average fragment sizes above 350 bp were sonicated using a Covaris M220 Focused-ultrasonicator™ (Covaris, Woburn, MA, USA) following the manufacturer's protocol to obtain an average fragment size of 350 bp. Dual-indexed libraries for Illumina® sequencing were prepared using the DNA NEBNext® Ultra™ II Library Prep Kit and the NEBNext® Multiplex Oligos for Illumina® (Dual Index Primers Set 1 and 2) from New England BioLabs® (Ipswich, MA, USA) following the manufacturer's instructions but at half the recommended volumes. The quality of the libraries was evaluated on the TapeStation using High Sensitivity D1000 ScreenTape and the libraries were quantified using a Quantus Fluorometer. The final average library size including the adapters was c. 500 bp. Afterwards, the samples were pooled (8–24 samples/reaction) and enriched with the Angiosperms-353 probes (Johnson et al., 2018) following the manufacturer's instructions (myProbes® Manual v4.01, Arbor Biosciences, Ann Arbor, MI, USA) setting the hybridization temperature to 65°C for 24 h. Final products were again run on the TapeStation to assess quality (i.e., average fragment size) so they could be pooled equimolarly for sequencing (48–96 samples/pool). After multiplexing library pools, sequencing was performed on an Illumina® MiSeq instrument (San Diego, CA, USA)—with v2 (300-cycles at 2 × 150 bp) or v3 (600-cycles at 2 × 300 bp) chemistry at Royal Botanic Gardens, Kew (Richmond, Surrey, UK)—or on an Illumina® HiSeq (San Diego, CA, USA)—at either Macrogen (Seoul, South Korea) or GENEWIZ® (Leipzig, Germany), producing 2 × 150 bp long reads.

Molecular work for the accessions enriched with the Cyperaceae-specific probes was carried out at The Morton Arboretum (Lisle, IL, USA) and the Pritzker Laboratory of the Field Museum of Natural History (Chicago, IL, USA). Genomic DNA was extracted from leaf tissue obtained from silica preserved samples (**Supplementary Table 1B**) using the QIAGEN DNeasy Plant Mini Kit following the manufacturer's protocols (QIAGEN, Valencia, CA, USA) or a modified CTAB protocol (Doyle and Doyle, 1987). Samples were sonicated to a target fragment size of 550 bp using a Covaris E220 Focused-ultrasonicator™ (Wohurn, MA, USA). Sequencing libraries were prepared using the Illumina® TruSeq Nano HT DNA kit (San Diego, CA, USA). DNA libraries were checked for quality using an Agilent Technologies 2100 Bioanalyzer (Santa Clara, CA, USA) and their concentration quantified using a Qubit 2.0 Fluorometer (Life Technologies, Grand Island, NY, USA). Indexed samples were pooled in approximately equal quantities and the pool was enriched using the custom Cyperaceae-specific probes (Villaverde et al. in review) following the manufacturer's protocols for myBaits® kit (v3), i.e., we hybridized at 65°C for 16 h. The paired-end libraries were sequenced in one run (including a total of 88 accessions; Villaverde et al. in review) on an Illumina MiSeq (2 × 300 bp; 600 cycle v3) at the Pritzker Laboratory.

Bioinformatics, Contig Assembly, and Multi-Sequence Alignment

Raw reads were trimmed to remove adapter sequences and to remove portions of low quality with Trimmomatic v.0.36 (Bolger et al., 2014) using the setting LEADING:20 TRAILING:20

SLIDINGWINDOW:4:20 MINLEN:50. HybPiper v1.3.1 (Johnson et al., 2016) was used with default settings to process the quality-checked, trimmed reads. Paired reads of samples enriched with two different kits independently (Angiosperms-353 and Cyperaceae-specific) were mapped to targets using BWA v0.7.17 (Li and Durbin, 2009) and their respective nucleotide (DNA) target file (**Supplementary Data Sheet 2**); additionally, we also used BLASTx (Altschul et al., 1990) when using the Angiosperms-353 target loci with amino acid (AA) sequences (**Supplementary Data Sheet 3**). Summary statistics such as the percentage of reads mapping were generated using SAMtools flagstat v1.8 (Li et al., 2009). Mapped reads were then assembled into contigs with SPAdes v3.13.1 (Bankevich et al., 2012). Subsequently, exonerate v2.2 (Slater and Birney, 2005) was used to align the assembled contigs to their associated target sequence and remove intronic regions. Only exon data were analyzed in the current study in order to directly compare the information content provided by the targeted loci. HybPiper flags potential paralogs when multiple contigs are discovered mapping well to a single reference sequence. The program uses coverage and identity to a reference to choose a "main" sequence and denotes the others as potential paralogs. All loci flagged as potential paralogs were removed from downstream analyses. A list of the potential paralogs is provided (**Supplementary Table 3**).

The paralog-filtered consensus sequences for each locus were used to generate eight different datasets (**Table 1**). This allowed us to investigate the phylogenetic informativeness of the data generated by the two kits separately and allowed us to test the mergeability of the data generated by both kits. Four unmerged datasets were created: (dataset 1) loci targeted with the Angiosperms-353 kit for the 37 C4 *Cyperus* clade accessions enriched with this kit plus *Cyperus kyllingiella* (C3 *Cyperus* grade) as outgroup; (dataset 2) loci targeted with the Cyperaceae-specific baits for the eight C4 *Cyperus* clade accessions enriched with this kit plus *S. pungens* as outgroup; (dataset 3) loci targeted with the Angiosperms-353 kit for a subset of eight C4 *Cyperus* clade accessions enriched with this kit; (dataset 4) loci targeted with the Cyperaceae-specific baits for the eight C4 *Cyperus* clade accessions enriched with this kit. Four merged datasets were assembled: (dataset 5) loci targeted with the Angiosperms-353 baits for all 46 *Cyperus* accessions; (dataset 6) loci targeted with the Cyperaceae-specific kit for all 46 *Cyperus* accessions; (dataset 7) all targeted loci for all 46 *Cyperus* accessions; and (dataset 8) the 57 overlapping loci targeted by both bait kits for all 46 *Cyperus* accessions (retrieved from the Angiosperms-353 data). The overlapping loci are listed in **Supplementary Table 4**. Contigs were aligned using MAFFT v7 (Katoh and Standley, 2013) with the "–auto" option. The number of parsimony informative sites were calculated for each contig alignment using AMAS (Borowiec, 2016).

Dataset 3 and dataset 4 were analyzed to account for the difference in sampling size when comparing the number of PIS retrieved across locus alignments for the two targeted sequencing kits. In these datasets, the eight C4 *Cyperus* clade accessions from dataset 2 (representing loci targeted with the Cyperaceae-specific baits for the accessions enriched with this kit) were compared with a taxonomically equivalent subset of eight accessions from dataset 1 (representing loci targeted with the Angiosperms-353 kit for the accessions enriched with this kit). The accessions selected for this subsampling are indicated by an asterisk in **Supplementary Tables 1** and **2**, and are represented respectively in dataset 3 and dataset 4 by 1) four species of the former segregate genus *Kyllinga* plus a closely related species to match five species of former segregate genus *Kyllinga* (two of which are represented by the same species in both datasets); 2) one species of the former segregate genus *Ascolepis* to match one species of the former segregate genus *Ascolepis*; 3) *C. esculentus* (represented by the same accession in both datasets); and 4) *C. papyrus* L. to match *Cyperus rotundus* L. which are closely related species.

TABLE 1 | The eight datasets analyzed in this study.

	Loci targeted by	Accessions enriched with	Accessions included
Unmerged datasets	Angiosperms-353	Angiosperms-353	37 C4 *Cyperus* clade + *Cyperus kyllingiella* as outgroup
	Cyperaceae-specific	Cyperaceae-specific	8 C4 *Cyperus* clade + *Schoenoplectus pungens* as outgroup
	Angiosperms-353	Angiosperms-353	8 C4 *Cyperus* clade
	Cyperaceae-specific	Cyperaceae-specific	8 C4 *Cyperus* clade
Merged datasets	Angiosperms-353	All accessions	46 *Cyperus* accessions
	Cyperaceae-specific	All accessions	46 *Cyperus* accessions
	All loci	All accessions	46 *Cyperus* accessions
	57 overlapping loci	All accessions	46 *Cyperus* accessions

Phylogenetic Inference

Trees were inferred using either a summary method that is statistically consistent under the Multiple Species Coalescent (MSC) (i.e., ASTRAL-III) or a total evidence approach, in which maximum likelihood (ML) inference was conducted on a concatenated matrix of all loci. Both approaches were used to analyze the eight different datasets described above (**Table 1**). For the summary approach under the MSC, individual gene trees were constructed using RAxML v8 (Stamatakis, 2014) applying GTRCAT and 200 bootstrap replicates followed by slow ML optimization with the "-f a" option. We then ran ASTRAL-III v5.5.11 (Zhang et al., 2018) to infer a species tree using "-t 2" to output quartet support values visualizing gene tree conflict. For the total evidence approach, phylogenetic inference of the targeted sequencing data was executed in IQ-TREE v1.6.10 (Nguyen et al., 2015) with 1,000 ultrafast bootstraps using the "- bb" and "-m TEST" options. To investigate gene tree *versus* species tree concordance, we calculated quartet distance between each individual gene tree and the concatenated total evidence tree obtained using the R package Quartet v1.0.2 (Sand et al., 2014; Smith, 2019), which yields a measure of the similarity of each gene tree *versus* the species tree based on shared four-taxon subtrees. We also calculated two measures of genealogical

concordance in our dataset, the gene concordance factor (gCF) and the site concordance factor (sCF), using the options "-gcf" and "-scf" in IQ-TREE v1.7beta (Nguyen et al., 2015; Minh et al., 2018). This approach provides a description of possible disagreement among loci and across sites.

RESULTS

Targeted Sequencing Kits and Data Quality

Summary statistics are available in **Supplementary Table 5**. When comparing the summary statistics between the two equally sized datasets 3 and 4 (**Supplementary Table 5C**), on average 324,655 (42,386–832,980) paired reads were produced for the accessions enriched with Angiosperms-353 probes *vs.* 199,961 (52,978–420,228) for the accessions enriched with the Cyperaceae-specific probes. Raw reads for all accessions are available from GenBank Sequence Read Archive (SRA) under BioProject numbers PRJNA553989 (*Cyperus* BioProject) and PRJNA553631 (*S. pungens—Carex* BioProject) (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA553989; http://www.ncbi.nlm.nih.gov/bioproject/PRJNA553631).

Universal *Versus* Family-Specific Probes

Recovery success and sequence length of the targeted loci, with both targeted sequencing kits, are provided in **Supplementary Table 2** and visualized in **Figure 2** and **Supplementary Images 1–3**. Percentage of recovery, i.e., the percentage of summed captured length of all target loci per individual divided by the summed mean length of all loci, was highest when running HybPiper for accessions enriched with the Cyperaceae-specific kit with its corresponding nucleotide target file (42.1%). We recovered on average 396 loci (324–476; **Supplementary Table 2C**) with the Cyperaceae-specific kit. For accessions enriched with the Angiosperms-353 kit, capture success was higher with the AA target file (33.23%) than with the DNA target file (21.7%). We recovered on average 215 loci (39–309; **Supplementary Table 2B**) with the AA target file and 162 loci (35–235) with the DNA target file. For data generated with the Angiosperms-353 kit, post-alignment length of contigs ranged from 87 to 3,103 bp long, with 751 bp mean length (**Table 2**; **Supplementary**

TABLE 2 | Length of the aligned contigs and number of parsimony informative sites (PIS) for data obtained after enrichment with the Angiosperms-353 and Cyperaceae-specific probes.

		Contig	PIS
Angiosperms-353	Mean	751	75
(38 accessions)	SD	438	65
Dataset 1	Min	87	0
	Max	3,103	439
	Total	**233,429**	**23,217**
Cyperaceae-specific	Mean	1,608	63
(9 accessions)	SD	830	59
Dataset 2	Min	93	0
	Max	7,527	479
	Total	**683,427**	**26,630**
Angiosperms-353	Mean	717	25
(subset of 8	SD	411	28
accessions)	Min	150	0
Dataset 3	Max	2,826	147
	Total	**221,564**	**7,613**
Cyperaceae-specific	Mean	1,471	50
(8 accessions)	SD	818	51
Dataset 4	Min	162	0
	Max	7,524	400
	Total	**667,945**	**22,767**

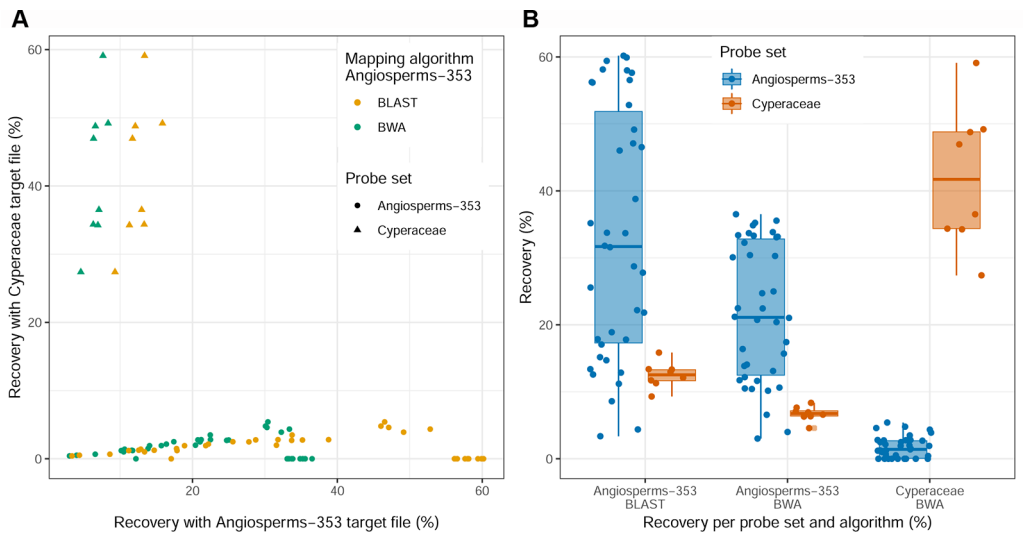

FIGURE 2 | Recovery success for the Angiosperms-353 kit *vs.* the Cyperaceae-specific kit. **(A)** Percentage recovery with Angiosperms-353 (AA) target file from the accessions enriched with the Angiosperms-353 kit *vs.* percentage recovery of the Cyperaceae-specific targets from the accessions enriched with the Cyperaceae-specific probes. **(B)** Summary of recovery success per kit and across kits.

Table 6A). For data generated with the Cyperaceae-specific kit, post-alignment length of contigs ranged from 93 to 7,527 bp long, with 1,608 bp mean length (**Table 2**; **Supplementary Table 6B**). In both cases, longer contigs had more Parsimony Informative Sites (PIS) (**Figure 3**).

We investigated three measures of resolution power when dealing with shallow-level phylogenetic relationships for the Angiosperms-353 kit *versus* the Cyperaceae-specific kit: 1) the proportion of gene trees that support the inferred species tree (under the MSC) (**Figure 3**; **Figures 4A, C**), 2) the disagreement among loci and across sites in the total evidence tree (**Figures 4B, D**), and 3) the number of PIS retrieved across locus alignments (**Figure 3**; **Supplementary Images 4** and **5**). Addressing the first measure, many nodes are well supported in the ASTRAL tree generated for dataset 1 (loci targeted with the Angiosperms-353 kit for the accessions enriched with this kit) (**Figure 4A**) having local posterior probability (LPP) values greater than 0.9. Addressing the second measure, many nodes are similarly well supported in the IQ-TREE tree generated with the same data (**Figure 4B**) with most nodes having bootstrap (BS) values greater than 90%. However, some of the branches that received low LPP or BS value support have quartet scores indicating gene tree conflict and/or have low gCF scores, which indicates that few gene trees support the grouping. These branches occur among some species of the *Cyperus margaritaceus-Cyperus niveus* complex (clade A) and in the main backbone of clade B, which represents species of C4 *Cyperus* s.s. and the 10 C4 segregate genera accepted by Goetghebeur (1998). In the phylogenetic hypotheses obtained for dataset 2 (loci targeted with the Cyperaceae-specific probes for the accessions enriched with this kit) (**Figures 4C, D**), most nodes are equally well supported having LPP values greater than 0.9 or BS values greater than 90%. Addressing the third measure, when comparing the probe sets in terms of PIS by comparing dataset 4 (loci targeted by the Cyperaceae-specific probes for the eight C4 *Cyperus* clade accessions enriched with this kit; **Supplementary Table 6D**) with the equally sized and taxonomically equivalent dataset 3 (loci targeted by the Angiosperms-353 probes for a subset of eight C4 *Cyperus* clade accessions enriched with this kit; **Supplementary Table 6C**), the former has an average contig length of 1,471 (162–7,524), while the latter has shorter average length of 717 (150–2,826). However, the relative number of PIS is the same at 0.03 PIS/bp. This is also shown in **Supplementary Image 4**. A comparison of the support provided by the gene trees for the species tree between the two datasets of eight accessions is provided in **Supplementary Image 5**.

Mergeability of Data Obtained With Different Targeted Sequencing Kits

Different percentages of recovery were obtained using the AA or the DNA target files of the Angiosperms-353 kit in the accessions enriched with the Cyperaceae-specific kit. With the AA target file, we were able to recover 12.5% of the total target size from accessions enriched with the Cyperaceae-specific kit (**Supplementary Table 2B**). This percentage decreases to 6.73% when using the DNA target file (**Supplementary Table 2A**). On average 44 loci (35–53) were retrieved from accessions enriched with the Cyperaceae-specific kit using the Angiosperms-353 AA target file, and 32 (21–41) with the DNA target file. Capture success was very low when targeting Cyperaceae-specific loci from accessions enriched with the Angiosperms-353 probes (1.7%) using the DNA target file; however, sequence data was still retrieved from an average of 37 loci (0–106; **Supplementary Table 2C**). This information is summarized in **Figure 2**.

We tested the mergeability of the data generated for all *Cyperus* samples produced after enrichment with the two

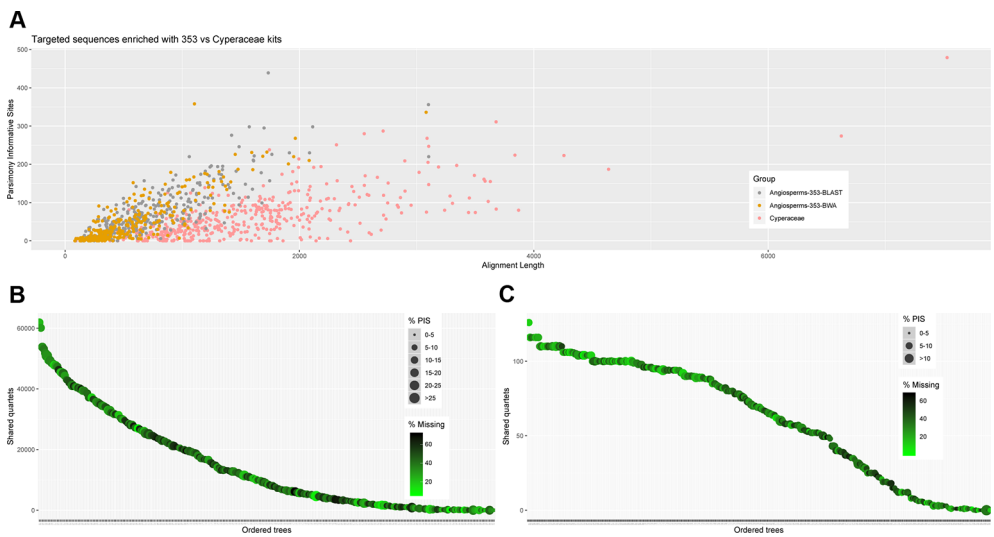

FIGURE 3 | (A) Scatter plot of aligned contig length *versus* number of parsimony informative sites. **(B)** Number of gene trees that support the inferred species tree for data generated with the Angiosperms-353 probes, and **(C)** for data generated with the Cyperaceae-specific kit.

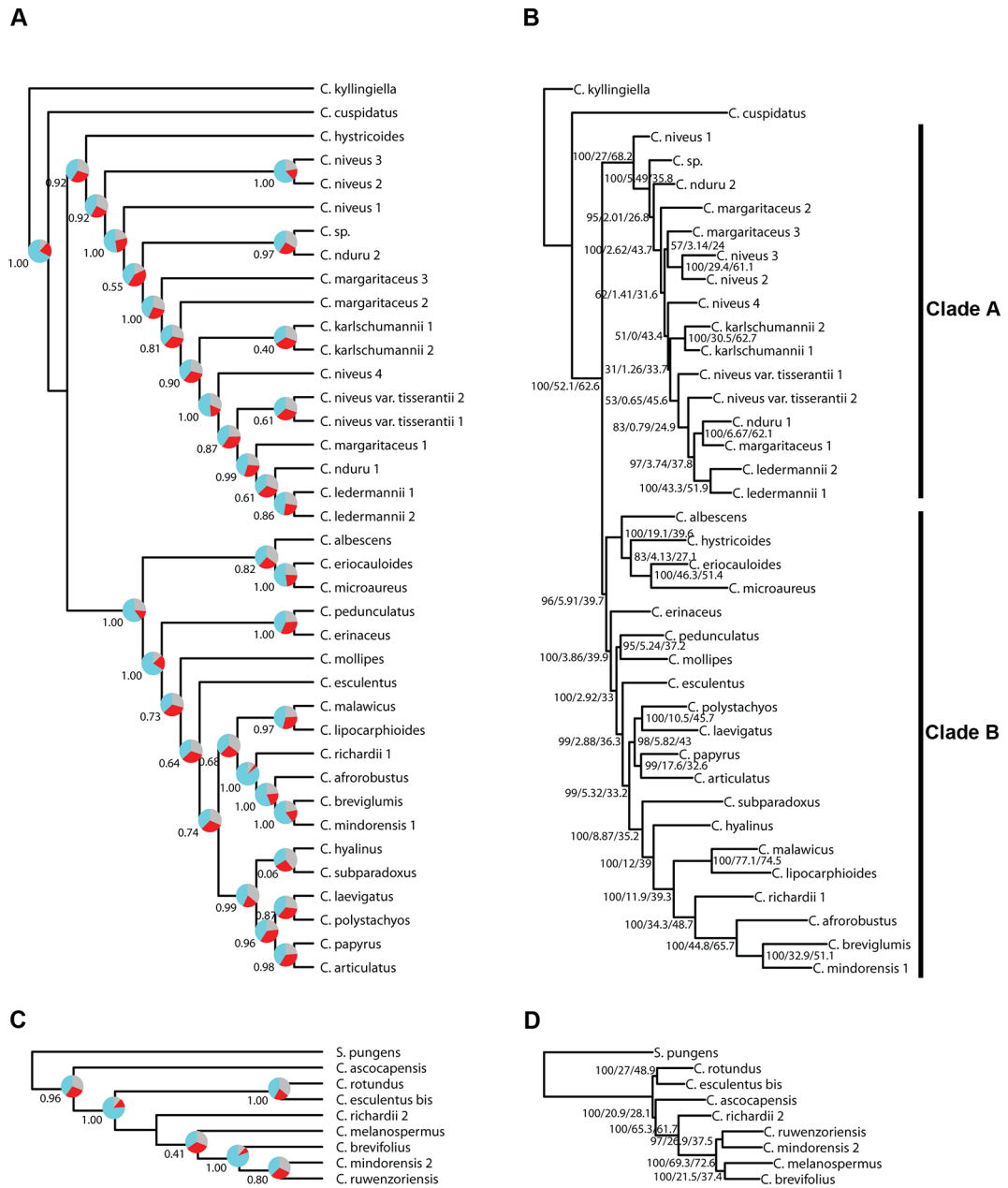

FIGURE 4 | Phylogenetic reconstructions using ASTRAL **(A, C)** and using IQ-TREE **(B, D)** of the relationships in the C4 *Cyperus* clade inferred for **(A, B)** dataset 1, i.e., 38 samples enriched with the Angiosperms-353 probes, and **(C, D)** dataset 2, i.e., 9 samples enriched with the Cyperaceae-specific probes. The ASTRAL trees show local posterior probability values and pie charts visualizing quartet support values at the nodes (blue = agreeing loci; red = disagreeing loci; gray = uninformative), while the IQ-TREE trees show BS/gCF/sCF values at the nodes.

targeted sequencing kits, by inferring trees using ASTRAL and IQ-TREE for four merged datasets. **Table 3** provides the length of the aligned contigs and number of PIS for the four merged datasets. The number of PIS is positively correlated with the size of the dataset (**Table 3**; **Supplementary Tables 6E–H**).

The amount of data retrieved targeting the Cyperaceae-specific loci from the data generated for all samples is larger when comparing the length of the alignments (**Table 3**; **Figure 5**) and the total number of PIS is also higher (c. 26.5% more PIS). The 57 overlapping loci are on average longer than those of most loci targeted by the Angiosperms-353 probes (1,206 *versus* 793) but shorter than the average for loci targeted with the Cyperaceae-specific kit (1,206 *versus* 1,458). These overlapping loci have a higher proportion of PIS per alignment in comparison with the other merged datasets (152 *versus* 74–85) (**Supplementary Image 6**; **Supplementary Tables 6E–H**).

TABLE 3 | Length of the aligned contigs and number of parsimony informative sites (PIS) for the four merged datasets: 1) aligned contigs of the loci targeted with the Angiosperm-353 probes, 2) aligned contigs of the loci targeted with the Cyperaceae-specific probes, 3) aligned contigs of all loci (without duplicating the overlapping genes), and 4) aligned contigs of the 57 overlapping loci targeted by both kits.

		Contig	PIS
Angiosperms-353	Mean	739	85
Dataset 5	SD	492	75
	Min	87	0
	Max	4,211	539
	Total	**259,321**	**27,656**
Cyperaceae-specific	Mean	1,458	76
Dataset 6	SD	805	81
	Min	153	0
	Max	7,520	554
	Total	**720,324**	**37,645**
All loci	Mean	1,192	74
Dataset 7	SD	775	74
	Min	87	0
	Max	7,520	554
	Total	**910,926**	**56,615**
57 overlapping loci	Mean	1,206	152
Dataset 8	SD	687	100
	Min	312	6
	Max	4,211	539
	Total	**68,719**	**8,686**

Although it includes much less data than the other analyzed datasets, the dataset of the 57 overlapping loci still includes a high number of PIS (8,686 PIS out of 68,719 bp or 0.12 PIS/bp; **Table 3**; **Supplementary Table 6H**).

Supplementary Table 4 lists the names of the overlapping loci for both targeted sequencing kits and a comparison of statistics between the recovery of these loci is provided in **Supplementary Table 6I**. The average contig length of the overlapping genes retrieved with the Angiosperms-353 AA target file is shorter than when they are retrieved with the Cyperaceae-specific DNA target file (1,206 *vs.* 1,482). However, the average number of PIS retrieved with the Angiosperms-353 AA target file is higher than when the data are retrieved with the Cyperaceae-specific DNA target file (152 *vs.* 126).

Resolving Relationships in the C4 *Cyperus* Clade

Topologies produced with both approaches for dataset 1 (loci targeted with the Angiosperms-353 kit for the accessions enriched with this kit; **Figures 4A, B**) are very similar, except for the position of *Cyperus hystricoides* (B. Nord.) Bauters, which is retrieved as sister to clade A in the ASTRAL analysis (**Figure 4A**) and as part of clade B in the IQ-TREE analysis (**Figure 4B**). Likewise, the trees generated with both approaches, using dataset 2 (loci targeted with the Cyperaceae-specific probes for the accessions enriched with this kit; **Figures 4C, D**), result in similar topologies, except for the placement of *Cyperus ascocapensis* Bauters.

The topologies in the ASTRAL trees resulting from the four merged datasets are very similar (**Figure 6**), with generally high levels of node support. As in **Figure 4**, the placement of *C. hystricoides* was unstable, being reconstructed as sister to clade A in the tree based on the loci targeted by the Angiosperms-353 kit (**Figure 6A**), and among the first branching lineages of clade B in the other analyses (**Figures 6B, D**). Other differences in topology were found among taxa of the *C. margaritaceus-C. niveus* complex (clade A) and in the backbone of clade B where node support is lower and quartet scores indicate higher gene tree discordance (**Figure 6**). The proportion of gene trees supporting the retrieved topology is similar in **Figures 6A–C**, but the ASTRAL analysis of dataset 8 (overlapping loci for all accessions), yields a tree with higher locus concordance at most nodes.

The analyses performed with IQ-TREE yielded topologies similar (**Figure 7**) to those retrieved with ASTRAL for the

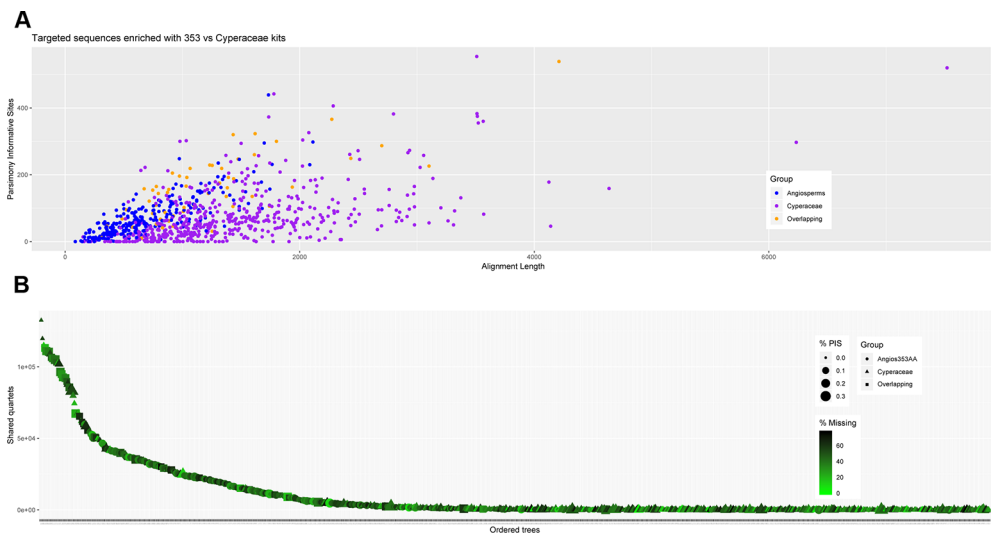

FIGURE 5 | Number of gene trees that support the inferred species tree **(A)**; and **(B)** scatter plot of aligned contig length *vs.* parsimony informative sites (PIS) for the data generated for all samples recovered targeting the Angiosperms-353 genes, the Cyperaceae-specific genes, and indicating the overlapping genes.

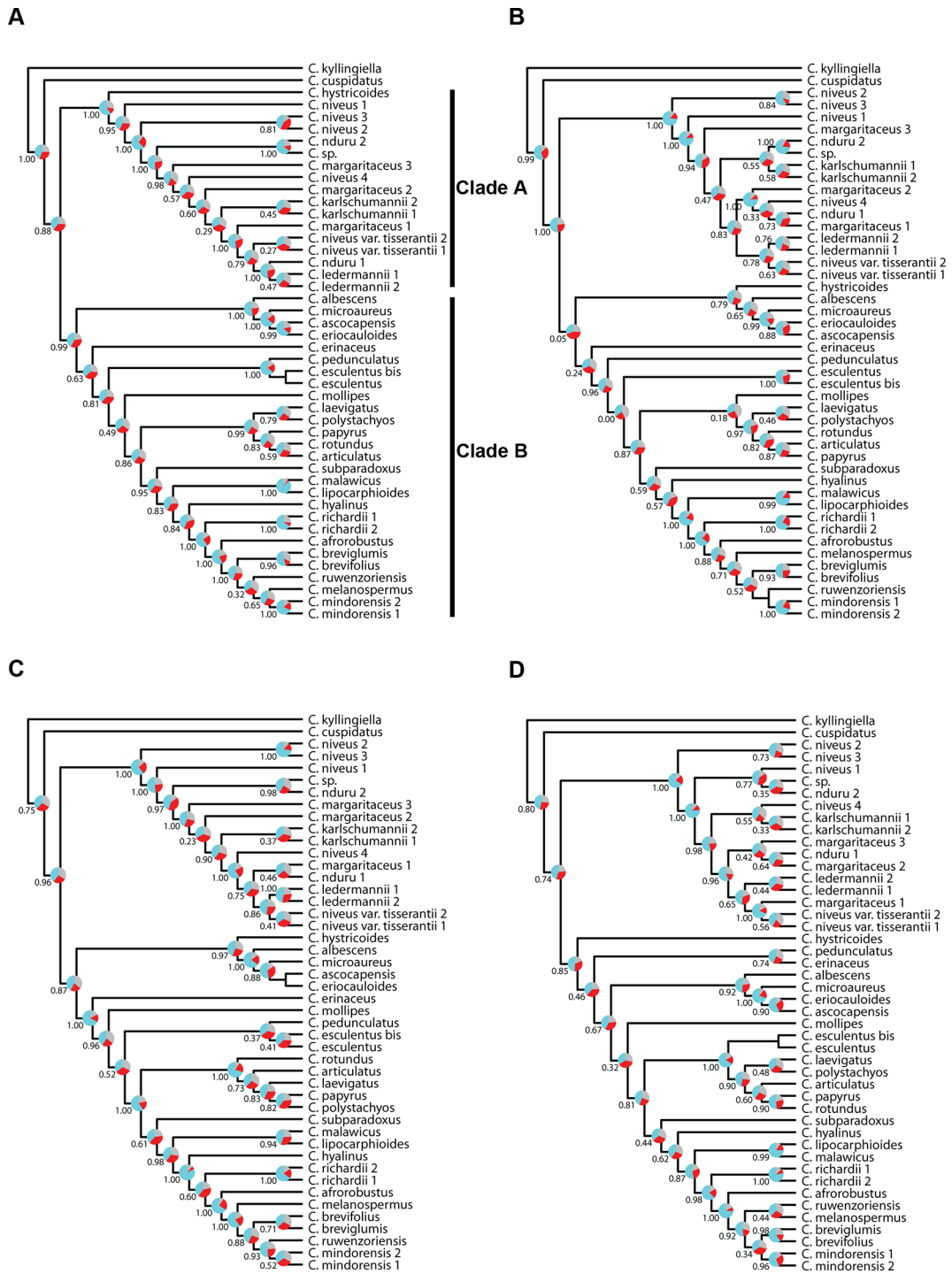

FIGURE 6 | Phylogenetic reconstructions using ASTRAL of the relationships in the C4 *Cyperus* clade inferred for all accessions from aligned contigs of **(A)** dataset 5, i.e., the loci targeted with the Angiosperms-353 probes, **(B)** dataset 6, i.e., the loci targeted with the Cyperaceae-specific kit, **(C)** dataset 7, i.e., all targeted loci, and **(D)** dataset 8, i.e., the overlapping loci targeted by both kits. The trees show local posterior probability values and pie charts visualizing quartet support values at the nodes (blue = agreeing loci; red = disagreeing loci; gray = uninformative).

respective datasets. For clade A, a morphologically homogeneous species complex in the C4 *Cyperus* clade, resolution, and support are comparable between the ASTRAL and IQ-TREE results (**Figures 6** and **7**). However, for clade B, a morphologically heterogeneous subset of the C4 *Cyperus* clade, the IQ-TREE analyses provided higher support for some nodes, although the IQ-TREE topology is often less well resolved (**Figures 6B, D** *vs.* **Figures 7B, D**). For dataset 6 (loci targeted with the Cyperaceae-

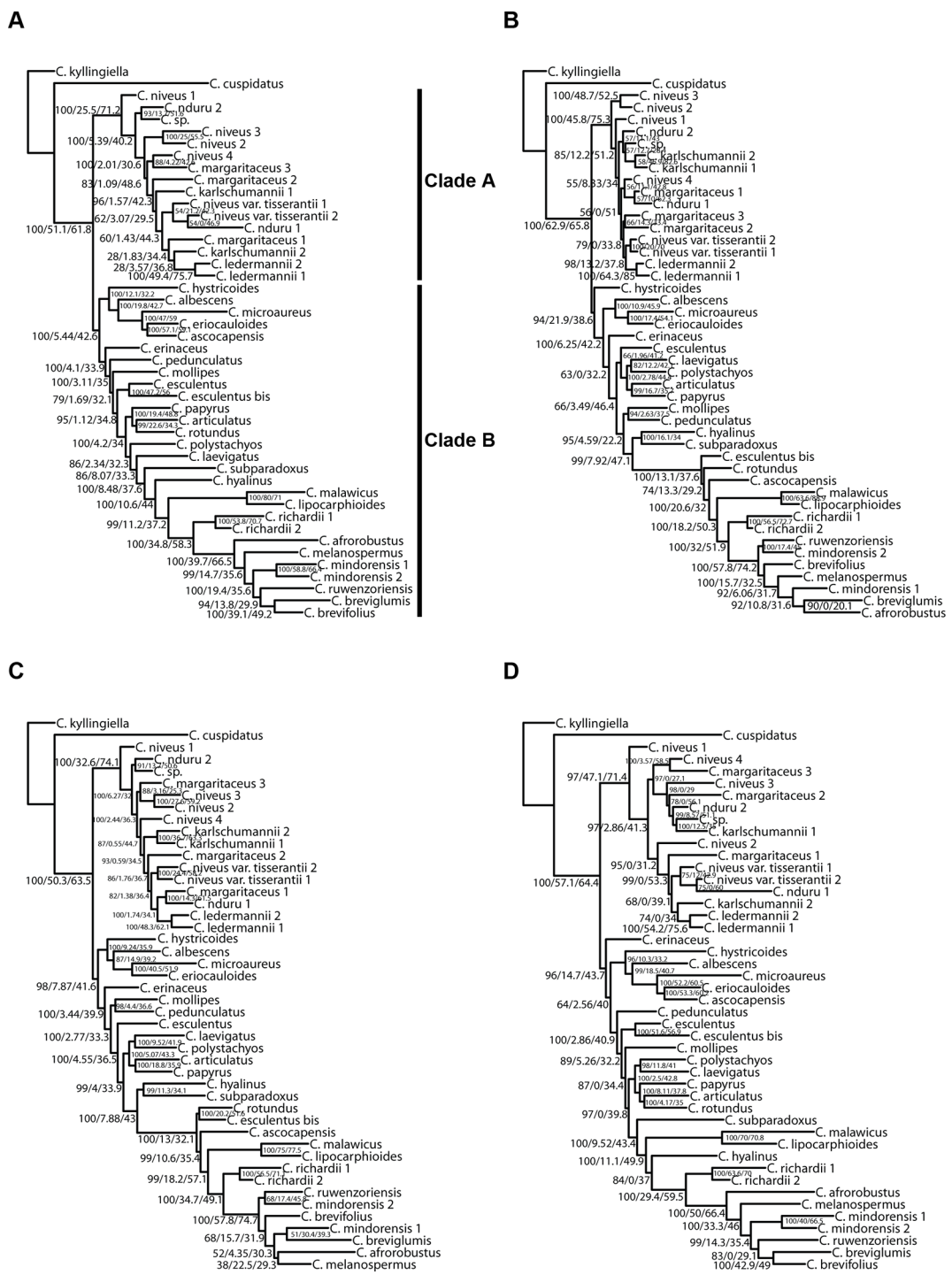

FIGURE 7 | Phylogenetic reconstructions using IQ-TREE of the relationships in the C4 *Cyperus* clade inferred for all accessions from aligned contigs of **(A)** dataset 5, i.e., the loci targeted with the Angiosperms-353 probes, **(B)** dataset 6, i.e., the loci targeted with the Cyperaceae-specific kit, **(C)** dataset 7, i.e., all targeted loci, and **(D)** dataset 8, i.e., the overlapping loci targeted by both kits. The trees show bootstrap/gene concordance factor/site concordance factor values at the nodes.

specific probes for all accessions) and dataset 7 (all loci for all accessions), ASTRAL, a summary species-tree method, seems the handle high levels of missing data better in that the obtained results retrieve conspecific accessions or closely related species as sister taxa (**Figures 6B, C**). On the other hand, the total evidence trees inferred from concatenated data matrices in IQ-TREE under ML does not accurately place several samples in clade B (**Figures 7B, C**), i.e., *C. esculentus* (same accession sequenced with both targeted sequencing kits) is not reconstructed as monophyletic, and neither is *C. mindorensis* (different accessions but the same species). Similarly, the IQ-TREE analysis of dataset 8 (overlapping loci for all accessions; **Figure 7D**), did not reconstruct *Cyperus ledermannii* (Kük.) S.S. Hooper, *C. niveus* var. *tisserantii* (Cherm.) Lye, *C. mindorensis*, and *C. richardii* as monophyletic. This issue is not found for taxa of clade B in the IQ-TREE ML analysis of dataset 5 (loci targeted with the Angiosperms-353 kit for all accessions), although placement of *Cyperus karlschumannii* C.B. Clarke and *C. niveus* var. *tisserantii* were not reconstructed as expected in clade A (**Figure 7A**).

When considering gene tree concordance for the analyses of dataset 5 (loci targeted with the Angiosperms-353 kit for all accessions), the monophyly of conspecific accessions or accessions of closely related species [e.g., *Cyperus lipocarphioides* (Kük.) Lye and *Cyperus malawicus* (J. Raynal) Lye] in clade B is supported by a high number of gene trees (**Figure 7A**). In clade A, only a few taxa are resolved as monophyletic (**Figure 7A**). In contrast, in the IQ-TREE results of dataset 6 (loci targeted with the Cyperaceae-specific probes for all accessions), and in the analysis of dataset 7 (all loci for all accessions), conspecific accessions in clade A tend to be retrieved as monophyletic and supported by a significant number of gene trees, while phylogenetic relationships between the taxa of clade B are not well resolved (**Figures 7B, C**). In the analyses based on dataset 8 (overlapping loci for all accessions), most nodes in clade B are supported by a proportion of gene trees, while many nodes in clade A have a gCF value of 0 (**Figure 7D**). This result contrasts with the higher locus concordance observed at most nodes in the ASTRAL analysis of this dataset (**Figure 6D**).

DISCUSSION

Data Quality: Herbarium *Versus* Tissue Preserved for Deoxyribonucleic Acid Extraction

Most tissue samples in this study (30 out of 47) were obtained from herbarium specimens. The remainder were from silica dried samples (17 samples). The high quality of the sequence capture across types illustrates the potential of targeted sequencing to generate genomic level data from fragmented DNA (Hart et al., 2016; McKain et al., 2018; Villaverde et al., 2018; Brewer et al., 2019). Differences in extraction and sequencing methods did not appear to influence capture success or recovered target length with respect to total length targeted (**Supplementary Table 2**).

Universal *Versus* Family-Specific Probes

Recovery of loci targeted with the Cyperaceae-specific probes from accessions enriched with this kit was higher than for loci targeted with the universal Angiosperms-353 kit (**Supplementary Table 2**). Although more data was retrieved with the Cyperaceae-specific kit (dataset 2) than with the Angiosperms-353 kit (dataset 1) (total length aligned contigs was 683,427 bp and 233,429 bp, respectively), the number of PIS was comparable (26,630 *vs.* 23,217, respectively), yielding a lower ratio of PIS per total length aligned for the Cyperaceae-specific kit (0.04 PIS/bp) than for the Angiosperms-353 kit (0.1 PIS/bp) (**Table 2**). However, when comparing the values between the equally sized and taxonomically equivalent datasets 3 and 4, a comparable ratio of PIS per total aligned contig length is obtained (both with 0.03 PIS/bp). This indicates that the power of both targeted sequencing kits (off-target reads notwithstanding) to resolve shallow-level relationships for the C4 *Cyperus* clade is quite similar (**Figure 3**; **Table 2**).

Our finding belies our expectation going into this study that our family-specific kit would recover more variable loci. A recent study (Kadlec et al., 2017) on the species-rich heather genus *Erica* concluded that data from markers, custom-designed using the MarkerMiner pipeline, deliver better results than those obtained using a more universal approach. In another recent study, Chau et al. (2018) developed a custom-designed targeted sequencing kit using data from two *Buddleja* species (Scrophulariaceae) from 1KP using a modified version of the Sondovac marker development pipeline (Schmickl et al., 2016). They compared these taxon-specific loci with three universal loci sets (a conserved ortholog set [COSII] by Wu et al., 2006, shared single-copy nuclear [APVO SSC] genes by Duarte et al., 2010; and pentatricopeptide repeat [PPR] genes by Yuan et al., 2009). Chau et al. (2018) conclude that targeted sequencing is an effective method for increasing resolution and support in phylogenetics compared to Sanger sequencing, and that universal target loci can be as effective as taxon-specific loci in terms of capture success and phylogenetic informativeness. Our results support the conclusions of Chau et al. (2018). The advantage of a universal off-the-shelf targeted sequencing kit like the Angiosperms-353 kit is that it opens the opportunity to use targeted sequencing in plant groups with few genomic resources (Chau et al., 2018). Furthermore, universal kits are attractive in terms of reduced cost and effort, because they generate data suitable for wider analyses across angiosperms and may be applied as a DNA barcoding tool (Blattner, 2016; Kadlec et al., 2017), and predesigned kits can often be purchased at a discount from the producers (https://arborbiosci.com/products/mybaits-plant-angiosperms/).

Mergeability of Data Obtained with Different Targeted Sequencing Kits

Villaverde et al. (in review) combined accessions enriched with the anchored phylogenomics probes (Buddenhagen et al., 2016; Léveillé-Bourret et al., 2018) in their analyses of *Carex* using the Cyperaceae-specific kit. Here, we go one step further: we combine accessions enriched with two different kits, and we

merge the data in both directions. Our study may be the first to do so, at least in angiosperms. As expected, sequence data recovery is higher when analyses are performed with the target file that matches the loci for which the samples were enriched (**Figure 2**). Nonetheless, even with limited overlap (57 genes) between targeted sequencing kits and a high level of missing data, data analyses under the MSC recover the same overall topology irrespective of dataset, with strong support. However, combined data analyses using a concatenated ML approach appear less robust to dataset differences and result in conspecific accessions not always being reconstructed as monophyletic when analyzing the merged datasets.

Resolving Relationships in the C4 *Cyperus* Clade

Most nodes are well supported in all analyses we conducted in the C4 *Cyperus* clade (**Figures 4**, **6**, and **7**), except for the branches near the backbone of clade B, as has been observed in earlier studies (e.g., Larridon et al., 2013; Bauters et al., 2014; Semmouri et al., 2019). These nodes show a higher gene tree discordance (based on ASTRAL quartet score and IQ-TREE gCF and sCF values), which likely resulted from an increased diversification rate (Spalink et al., 2016). Still the resolution and support retrieved in the backbone of the C4 *Cyperus* clade from targeted sequencing data is an important improvement over the polytomy obtained with Sanger sequencing results (e.g., Larridon et al., 2013).

The relationships retrieved in the C4 *Cyperus* clade, here investigated for the first time using phylogenomic data, largely match those obtained in previous studies (e.g., Larridon et al., 2011b; Larridon et al., 2013), with *Cyperus cuspidatus* Kunth sister to all other taxa in the C4 *Cyperus* clade. This species has an inflorescence of digitately clustered spikelets, which is characteristic of species in the C3 *Cyperus* grade and C4 *Cyperus* section *Amabilis* C.B. Clarke. Previously, sections in the C3 *Cyperus* grade + the C4 *Cyperus* section *Amabilis* were placed together in *Cyperus* subgenus *Pycnostachys* C.B. Clarke based on this shared inflorescence type *versus* the remaining sections in the C4 *Cyperus* clade, which are characterized by having spikes of spikelets. A notable difference with earlier studies is that remaining species in the C4 *Cyperus* clade form two well-supported clades (indicated as clade A and B; **Figures 4**, **6**, and **7**). One of the two groups of closely related species included in this study—i.e., the white-glumed *Cyperus* species or the *C. margaritaceus-C. niveus* complex (clade A)—is here reconstructed as sister to a clade (B) comprising the rest of the C4 *Cyperus* clade. Species of clade A had not been included in previous molecular studies. More research is needed to confirm that this is not a sampling artifact, however, although sampling in this study is limited, it adequately covers the range of morphological diversity observed in the C4 *Cyperus* clade, as it encompasses both species of C4 *Cyperus* s.s. (e.g., type species *C. esculentus* L.) and all 10 of the C4 segregate genera recognized by Goetghebeur (1998).

The position of *C. hystricoides* is unstable, being inferred either as sister to clade A or as part of the species-poor lineages in clade B (**Figures 4**, **6**, and **7**). This species was placed in *Rikliella* J. Raynal (Raynal, 1973) and later merged into *Lipocarpha* by Goetghebeur and Van den Borre (1989), who interpreted the inflorescence as a head of several spikes of spirally arranged single-flowered spikelets lacking a spikelet prophyll and glumes. However, an ontogenetic study (Bauters et al., 2014) showed that the inflorescence should be interpreted as a head of several spikelets with multiple spirally arranged flowers that have both a spikelet prophyll and glumes. With the new interpretation, the inflorescence type in *C. hystricoides* is similar to that of species previously placed in the C3 segregate genus *Kyllingiella* (now part of *Cyperus* sect. *Leucocephali*, incl. *Cyperus kyllingiella*; Larridon et al., 2011a), which is sister to the C4 *Cyperus* clade. This could provide morphological arguments for the placement of *C. hystricoides* among the species-poor lineages, away from the crown, of the C4 *Cyperus* clade.

Besides the *C. margaritaceus-C. niveus* complex, the other group of closely related species included in this study are seven species of *Cyperus* section *Kyllinga* (Rottb.) J. Kern (e.g., its type species *C. mindorensis*). Nodes within *Cyperus* sect. *Kyllinga* are well supported (**Figures 4**, **6**, and **7**), demonstrating the utility of the data obtained with both targeted sequencing kits to resolve low-level relationships in the C4 *Cyperus* clade. However, in the *C. margaritaceus-C. niveus* complex relationships between taxa are poorly supported although most morphologically defined taxa are retrieved as monophyletic, at least in the ASTRAL analyses (**Figures 4**, **6**, and **7**).

The results confirm the close relationship between *Cyperus laevigatus* (placed in the former segregate genus *Juncellus* C.B. Clarke) and *Cyperus polystachyos* (type species of the former segregate genus *Pycreus*) found in previous studies (e.g., Larridon et al., 2013; Semmouri et al., 2019). As in C3 *Cyperus* and C4 *Cyperus* s.s., the species previously placed in *Juncellus* and *Pycreus* have spikelets with multiple distichously arranged glumes each bearing a flower. However, in contrast to *Cyperus* s.s. with trigonous nutlets, *Juncellus* was recognized by dorsiventrally flattened nutlets, while in *Pycreus* nutlets are laterally compressed. In Cyperoideae, the development of the gynoecium from an annular primordium facilitates the shift in localization of stigma primordia (Vrijdaghs et al., 2011). Together with the decoupled development of the ovary and ovule (Reynders et al., 2012), this enables shifts between trigonous and dorsiventrally and laterally flattened nutlets in related species.

Our study is also the first to include all four species of the former segregate genus *Alinula* (*C. lipocarphioides*, *C. malawicus*, *Cyperus microaureus* Lye, *Cyperus subparadoxus* Kük.; **Figures 4**, **6**, and **7**). Earlier efforts to include all species in a Sanger sequencing study had failed due to degraded DNA extracted from herbarium specimens. This illustrates the advantage of targeted sequencing over Sanger sequencing for degraded DNA. *Alinula* sensu Goetghebeur (1998) is clearly polyphyletic (three groups). The first species to be published in *Alinula* was *C. lipocarphioides* by Raynal (1977) when he described the new genus. Later, another species, *C. malawicus*, was suggested to be its closest relative (Haines and Lye, 1983; Goetghebeur, 1986; Goetghebeur and Vorster, 1988). Our results confirm this close relationship. The species *Cyperus microaureus* was originally described in the segregate genus *Ascolepis*, but Goetghebeur (1977) relegated it to its own monotypic genus *Marisculus* Goetgh. because some of its

inflorescence and spikelet characteristics are peculiar. Later, it was placed in *Alinula* (Goetghebeur and Vorster, 1988). In our results, the species appears sister to *Ascolepis* [represented by *C. ascocapensis* and *Cyperus eriocauloides* (Steud.) Bauters]. In his doctoral thesis, Vorster (1978) placed the fourth species, *Cyperus subparadoxus*, in a monotypic genus *Pseudolipocarpha* (not validly published) before moving it to *Alinula* (Goetghebeur and Vorster, 1988. It is here retrieved as a lineage separate from the other species formerly placed in *Alinula*.

CONCLUSION

We show the utility of two targeted sequencing kits, the universal Angiosperms-353 kit and a Cyperaceae-specific kit, in resolving relationships in a fast-evolving and taxonomically complex plant lineage, i.e., the C4 *Cyperus* clade. The probes from both kits work well with the often-degraded DNA-template obtained from herbarium material and allow the resolution of long-standing questions in Cyperaceae systematics (e.g., concerning the former segregate genus *Alinula*), where Sanger sequencing was previously either unsuccessful or provided no resolution. Generally, high support is retrieved using data of either or both kits, but some issues remain for the shortest branches where either significant conflict in gene trees or lack of signal occurs as shown by quartet scores, and gene and sCF. Potentially, adding off-target flanking regions and retrieving off-target high-copy sequence data such as the plastid genome, may provide added resolution. Our results demonstrate that data generated with a family-specific kit do not necessarily have more power than those obtained with a universal kit, at least in the C4 *Cyperus* clade, but that data generated with different targeted sequencing kits can often be merged for downstream analyses. Moreover, our study contributes to the growing consensus that targeted sequencing data are a powerful tool in resolving rapid radiations. We encourage ongoing studies to use targeted sequencing in lieu of Sanger sequencing to investigate the evolutionary history of Cyperaceae. The short-term costs in the lab will surely be mediated by long-term savings, as data can be repurposed for population genetics and phylogenetics with no return to the lab to sequence *just one more locus*.

DATA AVAILABILITY STATEMENT

The data generated for this study can be found in Genbank SRA under Bioproject numbers PRJNA553989 (Cyperus Bioproject), PRJEB35281 (Cyperus Baits Bioproject) and PRJNA553631 (Schoenoplectus pungens – Carex Bioproject).

AUTHOR CONTRIBUTIONS

IL and TV contributed equally as first authors. IL, TV, and AZ conceived the project design. IL and MX performed the sampling. IL, TV, LP, GB, NE, IF, MH, EM, and OM were responsible for generating the sequence data. IL, TV, AZ, and LP conducted the bioinformatic and molecular evolutionary analyses and wrote the manuscript. WB, FF, and AH supervised the research. All authors read and commented on the manuscript.

FUNDING

IL is supported by the B.A. Krukoff Fund for the Study of African Botany and a pilot study grant from the Royal Botanic Gardens, Kew. Phylogenomic work at Kew was funded by grants from the Calleva Foundation, the Sackler Trust and the Garfield Weston Foundation. The research of TV is supported by the Spanish Ministry of Economy and Competitiveness (project CGL2016-77401-P). The research of EM was supported by the Spanish Government, through Juan de la Cierva-Formación contract (FJCI-2017-32314). Carex NSF grant (NSF-DEB award #1255901) supported the lab work and sequencing at the Morton Arboretum. We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI).

ACKNOWLEDGMENTS

We would like to thank Sidonie Bellot for her help in visualizing the ASTRAL output and for always being happy to respond to lab and analysis related questions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01655/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Angiosperm Phylogeny Group (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

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* 3, e3376. doi: 10.1371/journal.pone.0003376

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.20120021

Bauckhage, C. (2015). *Numpy/scipy recipes for data science: k-medoids clustering* (Published on ResearchGate). doi: 10.13140/RG.2.1.46020564

Bauters, K., Larridon, I., Reynders, M., Huygh, W., Asselman, P., Vrijdaghs, A., et al. (2014). A new classification for *Lipocarpha* and *Volkiella* as infrageneric taxa of *Cyperus* s.l. (Cypereae, Cyperoideae, Cyperaceae): insights from species tree reconstruction supplemented with morphological and floral developmental data. *Phytotaxa* 166, 1–32. doi: 10.11646/phytotaxa.166.1.1

Beck, J. B., Alexander, P. J., Applhin, L., Al-Shehbaz, I. A., Rushworth, C., Bailey, C. D., et al. (2012). Does hybridization drive the transition to asexuality in diploid Boechera? *Evol.* 66, 985–995. doi: 10.1111/j.1558-5646.2011.01507.x

Blattner, F. R.. (2016). TOPO6: a nuclear single-copy gene for plant phylogenetic inference. *Plant Syst. Evol.* 302, 239–244. doi: 10.1007/s00606-015-1259-1

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* 4, e1660. doi: 10.7717/peerj1660

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

Buddenhagen, C., Lemmon, A. R., Lemmon, E. M., Bruhl, J. J., Cappa, J., Clement, W. L., et al. (2016). Anchored phylogenomics of angiosperms I: assessing the robustness of phylogenetic estimates. *BioRxiv [Preprint]*, 1–61. doi: 10.1101/086298

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

Chase, M. W., Soltis, D. E., Olmstead, R. G., Morgan, D., Les, D. H., Mishler, B. D., et al. (1993). Phylogenetics of Seed Plants: An Analysis of Nucleotide Sequences from the Plastid Gene rbcL. *Ann. Missouri Bot. Garden* 80, 528–548 & 550–580. doi: 10.2307/2399846

Chau, J. H., Rahfeldt, W. A., and Olmstead, R. G. (2018). Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Appl. Plant Sci.* 6, e1032. doi: 10.1002/aps3.1032

Couvreur, T. L. P., Helmstetter, A. J., Koenen, E. J. M., Bethune, K., Brandão, R. D., Little, S. A., et al. (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

Deschamps, S., Llaca, V., and May, G. D. (2012). Genotyping-by-Sequencing in Plants. *Biol.* 1, 460–483. doi: 10.3390/biology1030460

Dodsworth, S., Pokorny, S., Johnson, M. G., Kim, J. T., Maurin, O., Wickett, N. J., et al. (2019). Hyb-Seq for Flowering Plant Systematics. *Trends Plant Sci.* 24, 887–891. doi: 10.1016/j.tplants.2019.07.011

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, J. C., 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

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6, e19379. doi: 10.1371/journal.pone.0019379

Escudero, M., and Hipp., A. L. (2013). Shifts in diversification rates and clade ages explain species richness in higher-level sedge taxa (Cyperaceae). *Am. J. Bot.* 100, 2403–2411. doi: 10.3732/ajb.1300162

Escudero, M., Eaton, D. A. R., Hahn, M., and Hipp, A. L. (2014). Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in *Carex* (Cyperaceae). *Mol. Phyl. Evol.* 79, 359–367. doi: 10.1016/j.ympev.2014.06.026

Faircloth, B. C., Alda, F., Hoekzema, K., Burns, M. D., Oliveira, C., Albert, J. S., et al. (2018). A target enrichment bait set for studying relationships among ostariophysan fishes. *BioRxiv [Preprint]*, 1–53. doi: 10.1101/432583

Fishbein, M., Hibsch-Jetter, C., Soltis, D. E., and Hufford, L. (2001). Phylogeny of saxifragales (Angiosperms, Eudicots): analysis of a rapid, ancient radiation. *Syst. Biol.* 50, 817–847. doi: 10.1080/106351501753462821

Goetghebeur, P., and Van den Borre, A. (1989). Studies in Cyperaceae 8. A revision of *Lipocarpha*, including *Hemicarpha* and *Rikliella*. *Wageningen Agr. Univ. Pap.* 89, 1–87. http://edepot.wur.nl/283696.

Goetghebeur, P. (1977). Studies in Cyperaceae 1. Taxonomic notes on *Ascolepis* and *Marisculus*, a new genus of the tribe Cypereae. *Bull. Nat. Plantentuin Belg.* 47, 435–447. doi: 10.2307/3667910

Goetghebeur, P. (1986). "Genera Cyperacearum," in *Een bijdrage tot de kennis van morfologie, systematiek en fylogenese van de Cyperaceae. PhD thesis* (Gent: Ghent University), 1164.

Goetghebeur, P., and Vorster, P. (1988). Studies in Cyperaceae 7. The genus Alinula J. Raynal: A reappraisal. *Bull. Jard. Bot. Natl. Belg.* 58, 457–465. doi: 10.2307/3668298

Goetghebeur, P. (1998). "Cyperaceae," in *The families and genera of vascular plants 4*. Ed. K. Kubitzki (Berlin: Springer-Verlag), 141–190.

Haines, R., and Lye, K. A. (1983). *The sedges and rushes of East Africa*. (Nairobi: East African Natural History Society), p. 404.

Hart, M. L., Forrest, L. L., Nicholls, J. A., and Kidner, C. A. (2016). Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 65, 1081–1092. doi: 10.12705/655.9

Hinchliff, C. E., and Roalson, E. H. (2013). Using supermatrices for phylogenetic inquiry: an example using the sedges. *Syst. Biol.* 62, 205–219. doi: 10.1093/sysbio/sys088

Huygh, W., Larridon, I., Reynders, M., Muasya, A. M., Govaerts, R., Simpson, D. A., et al. (2010). Nomenclature and typification of names of genera and subdivisions of genera in Cypereae (Cyperaceae): 1. Names of genera in the *Cyperus* clade. *Taxon* 59, 1883–1890. doi: 10.1002/tax.596021

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. W., 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. (2018). 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

Jung, J., and Choi, H.-K. (2012). Recognition of two major clades and early diverged groups within the subfamily Cyperoideae (Cyperaceae) including Korean sedges. *J. Plant Res.* 126, 335–349. doi: 10.1007/s10265-012-0534-2

Kadlec, M., Bellstedt, D. U., Le Maitre, N. C., and Pirie, M. D. (2017). Targeted NGS for species level phylogenomics: "made to measure" or "one size fits all"? *PeerJ* 5, e3569. doi: 10.7717/peerj3569

Kates, H. R., Johnson, M. G., Gardner, E. M., Zerega, N. J. C., and Wickett, N. J. (2018). Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. *Am. J. Bot.* 105, 404–416. doi: 10.1002/ajb21068

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

Kipkemboi, J., and van Dam, A. A. (2018). "Papyrus Wetlands," in *The Wetland Book: II: Distribution, Description, and Conservation*. Eds. C. M. Finlayson, G. R. Milton, R. C. Prentice and N. C. Davidson (Dordrecht: Springer Netherlands), 183–197. doi: 10.1007/978-94-007-4001-3_218

Larridon, I., Reynders, M., Huygh, W., Bauters, K., Vrijdaghs, A., Leroux, O., et al. (2011a). Taxonomic changes in C3 *Cyperus* (Cyperaceae) supported by molecular phylogenetic data, morphology, embryology, ontogeny and anatomy. *Plant Ecol. Evol.* 144, 327–356. doi: 10.5091/plecevo.2011.653

Larridon, I., Reynders, M., Huygh, W., Bauters, K., Van de Putte, K., Muasya, A. M., et al. (2011b). Affinities in C3 *Cyperus* lineages (Cyperaceae) revealed using molecular phylogenetic data and carbon isotope analysis. *Bot. J. Linn. Soc* 167, 19–46. doi: 10.1111/j.1095-8339.2011.01160.x

Larridon, I., Reynders, M., Huygh, W., Muasya, A. M., Govaerts, R., Simpson, D. A., et al. (2011c). Nomenclature and typification of names of genera and subdivisions of genera in Cypereae (Cyperaceae): 2. Names of subdivisions of *Cyperus*. *Taxon* 60, 868–884. doi: 10.1002/tax.603021

Larridon, I., Bauters, K., Reynders, M., Huygh, W., Muasya, A. M., Simpson, D. A., et al. (2013). Towards a new classification of the giant paraphyletic genus *Cyperus* (Cyperaceae): phylogenetic relationships and generic delimitation in C4 *Cyperus*. *Bot. J. Linn. Soc* 172, 106–126. doi: 10.1111/boj.12020

Larridon, I., Bauters, K., Huygh, W., Reynders, M., and Goetghebeur, P. (2014). Taxonomic changes in C4 *Cyperus* (Cypereae, Cyperoideae, Cyperaceae): combining the sedge genera *Ascolepis*, *Kyllinga* and *Pycreus* into *Cyperus* s.l. *Phytotaxa* 166, 33–48. doi: 10.11646/phytotaxa.166.1.2

Lemmon, A. R., Emme, S. A., and Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* 61, 727–744. doi: 10.1093/sysbio/sys049

Léveillé-Bourret, É., Starr, J. R., Ford, B. A., Lemmon, E. M., and Lemmon, A. R. (2018). Resolving rapid radiations within angiosperm families using anchored phylogenomics. *Syst. Biol.* 67, 94–112. doi: 10.1093/sysbio/syx050

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liu, Y., Johnson, M. G., Cox, C. J., Medina, R., Devos, N., Vanderpoorten, A., et al. (2019). Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. *Nat. Commun.* 10, 1485. doi: 10.1038/s41467-019-09454-w

Maguilla, E., Escudero, M., Hipp, A. L., and Luceño, M. (2017). Allopatric speciation despite historical gene flow: Divergence and hybridization in *Carex furva* and *C. lucennoiberica* (Cyperaceae) inferred from plastid and nuclear RAD-seq data. *Mol. Ecol.* 26, 5646–5662. doi: 10.1111/mec.14253

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., et al. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience* 3, 10–17. doi: 10.1186/2047-217X-3-17

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/aps31038

Minh, B. Q., Hahn, M., and Lanfear, R. (2018). New methods to calculate concordance factors for phylogenomic datasets. *BioRXiv [Preprint]*, 1–7. doi: 10.1101/487801

Mitchell, N., Lewis, P. O., Lemmon, E. M., Lemmon, A. R., and Holsinger, K. E. (2017). Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. *Am. J. Bot.* 104, 102–115. doi: 10.3732/ajb.1600227

Muasya, A. M., Simpson, D. A., and Chase, M. W. (2002). Phylogenetic relationships in *Cyperus* L. s.l. (Cyperaceae) inferred from plastid DNA sequence data. *Bot. J. Linn. Soc* 138, 145–153. doi: 10.1046/j.1095-8339.2002.138002145.x

Muasya, A. M., Simpson, D. A., Verboom, G. A., Goetghebeur, P., Naczi, R. F. C., Chase, M. W., et al. (2009). Phylogeny of Cyperaceae based on DNA sequence data: current progress and future prospects. *Bot. Rev.* 75, 2–21. doi: 10.1007/s12229-008-9019-3

Nabhan, A. R., and Sarkar, I. N. (2012). The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief Bioinform.* 13, 122–134. doi: 10.1093/bib/bbr014

Nguyen, L.-T., Schmidt, H. A., van 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

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

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, T. J., Manuel, M., Wörheide, Gert, et al. (2011). Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PloS Biol.* 9, e1000602. doi: 10.1371/journal.pbio.1000602

Raynal, J.. (1973). Cyperological notes 19. Contribution to the classification of the Cyperoideae subfamily. *Adansonia Ser.* 2, 13 (2), 145–171.

Reynders, M., Huygh, W., Larridon, I., Muasya, A. M., Govaerts, R., Simpson, D. A., et al. (2011). Nomenclature and typification of names of genera and subdivisions of genera in the Cypereae (Cyperaceae): 3. Names in segregate genera of *Cyperus*. *Taxon* 60, 885–895. doi: 10.1002/tax.603022

Reynders, M., Vrijdaghs, A., Larridon, I., Huygh, W., Leroux, O., Muasya, A. M., et al. (2012). Gynoecial anatomy and development in Cyperoideae (Cyperaceae, Poales): congenital fusion of carpels facilitates evolutionary modifications in pistil structure. *Plant Ecol. Evol.* 145, 96–125. doi: 10.5091/plecevo.2012.675

Roalson, E. H., Prata, A. P., Mesterházy, A., Chase, M. W., Simpson, D. A., Thomas, W. W., et al. (2019). A broader circumscription of *Bulbostylis*, to include *Nemum* (Abildgaardieae: Cyperaceae). *Phytotaxa* 395, 199–208. doi: 10.11646/phytotaxa.395.3.4

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

Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., and Pedersen, C. N. S. (2014). tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics* 30, 2079–2080. doi: 10.1093/bioinformatics/btu157

Schmickl, R., Liston, A., Zeisek, V., Oberlander, K., Weitemier, K., Straub, S. C. K., et al. (2016). Phylogenetic marker development for target enrichment from transcriptomic and genome skim data: The pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Mol. Ecol. Resour.* 16, 1124–1135. doi: 10.1111/1755-0998.12487

Semmouri, I., Bauters, K., Léveillé-Bourret, É., Starr, J. R., Goetghebeur, P., and Larridon, I. (2019). The phylogeny and systematics of Cyperaceae, the evolution and importance of embryo morphology. *Bot. Rev.* 85, 1–39. doi: 10.1007/s12229-018-9202-0

Simpson, D. A., and Inglis, C. A. (2001). Cyperaceae of economic, ethnobotanical and horticultural importance: a checklist. *Kew Bull.* 56, 257–360. doi: 10.2307/4110962

Simpson, D. A., Muasya, A. M., Alves, M., Bruhl, J. J., Dhooge, S., Chase, M. W., et al. (2007). Phylogeny of Cyperaceae based on DNA sequence data—a new *rbcL* analysis. *Aliso* 23, 72–83. doi: 10.5642/aliso.20072301.09

Slater, G. S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* 6, 31. doi: 10.1186/1471-2105-6-31

Smith, M. R. (2019). Quartet: comparison of phylogenetic trees using quartet and bipartition measures. doi: 10.5281/zenodo.2536318

Snak, C., Vatanparast, M., Christian, S., Lewis, G. P., Lavin, M., Kajita, T., et al. (2016). A dated phylogeny of the papilionoid legume genus Canavalia reveals recent diversification by a pantropical liana lineage. *Mol. Phyl. Evol.* 98, 133–146. doi: 10.1016/j.ympev.2016.02.001

Spalink, D., Drew, B. T., Pace, M. C., Zaborsky, J. G., Starr, J. R., Cameron, K. M., et al. (2016). Biogeography of the cosmopolitan sedges (Cyperaceae) and the area-richness correlation in plants. *J. Biogeogr.* 43, 1893–1904. doi: 10.1111/jbi.12802

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, J. D., Rogers, W. L., Heyduk, K., Cruse-Sanders, J. M., Determann, R. O., Glenn, T. C., et al. (2015). Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. *Mol. Phyl. Evol.* 85, 76–87. doi: 10.1016/j.ympev.2015.01.015

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

Villaverde, T., Jiménez-Mejías, P., Luceño, M., Roalson, E. H., Hipp, A. L.the Global Carex Group (in review). A new classification of *Carex* subgenera supported by a Hyb-Seq backbone phylogeny. *Bot. J. Linn. Soc.*

Vorster, P. J. (1978). *Revision of the taxonomy of Mariscus Vahl and related genera in southern Africa. PhD thesis* (Pretoria: University of Pretoria), 384.

Vrijdaghs, A., Reynders, M., Muasya, A. M., Larridon, I., Goetghebeur, P., and Smets, E. (2011). Morphology and development of spikelets and flowers in Cyperus and Pycreus (Cyperaceae). *Plant Ecol. Evol.* 144, 44–63. doi: 10.5091/plecevo.2011.436

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

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

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. *PNAS* 111, E4859–E4868. doi: 10.1073/pnas.1323926111

Wu, F., Mueller, L. A., Crouzillat, D., Pétiard, V., and Tanksley, S. D. (2006). Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative evolutionary and systematic studies: A test case in the euasterid plant clade. *Genetics* 174, 1407–1420. doi: 10.1534/genetics.106.062455

Yuan, Y.-W., Liu, C., Marx, H. E., and Olmstead, R. G. (2009). The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. *New Phytol.* 182, 272–283. doi: 10.1111/j.1469-8137.2008.02739.x

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

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: 20 March 2020
doi: 10.3389/fpls.2020.00258

Reconstructing the Complex Evolutionary History of the Papuasian *Schefflera* Radiation Through Herbariomics

Zhi Qiang Shee[1,2], David G. Frodin[1†], Rodrigo Cámara-Leret[1,3,4] and Lisa Pokorny[1,5,6]*

[1] Royal Botanic Gardens, Kew, Richmond, United Kingdom, [2] Singapore Botanic Gardens, Singapore, Singapore, [3] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, [4] Bren School of Environmental Science and Management, University of California, Santa Barbara, Santa Barbara, CA, United States, [5] Centre for Plant Biotechnology and Genomics (CBGP UPM-INIA), Madrid, Spain, [6] Real Jardín Botánico (RJB-CSIC), Madrid, Spain

OPEN ACCESS

Edited by:
Thomas L. P. Couvreur, IRD UMR 232 Diversité, Adaptation, Développement des Plantes (DIADE), France

Reviewed by:
Andrew James Helmstetter, IRD UMR 232 Diversité, Adaptation, Développement des Plantes (DIADE), France
Laura Lowe Forrest, Royal Botanic Garden Edinburgh, United Kingdom

***Correspondence:**
Lisa Pokorny
pokorny@rjb.csic.es

† Deceased

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

Received: *03 September 2019*
Accepted: *19 February 2020*
Published: *20 March 2020*

Citation:
Shee ZQ, Frodin DG, 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

With its large proportion of endemic taxa, complex geological past, and location at the confluence of the highly diverse Malesian and Australian floristic regions, Papuasia – the floristic region comprising the Bismarck Archipelago, New Guinea, and the Solomon Islands – represents an ideal natural experiment in plant biogeography. However, scattered knowledge of its flora and limited representation in herbaria have hindered our understanding of the drivers of its diversity. Focusing on the woody angiosperm genus *Schefflera* (Araliaceae), we ask whether its morphologically defined infrageneric groupings are monophyletic, when these lineages diverged, and where (within Papuasia or elsewhere) they diversified. To address these questions, we use a high-throughput sequencing approach (Hyb-Seq) which combines target capture (with an angiosperm-wide bait kit targeting 353 single-copy nuclear loci) and genome shotgun sequencing (which allows retrieval of regions in high-copy number, e.g., organellar DNA) of historical herbarium collections. To reconstruct the evolutionary history of the genus we reconstruct molecular phylogenies with Bayesian inference, maximum likelihood, and pseudo-coalescent approaches, and co-estimate divergence times and ancestral areas in a Bayesian framework. We find strong support for most infrageneric morphological groupings, as currently circumscribed, and we show the efficacy of the Angiosperms-353 probe kit in resolving both deep and shallow phylogenetic relationships. We infer a sequence of colonization to explain the present-day distribution of *Schefflera* in Papuasia: from the Sunda Shelf, *Schefflera* arrived to the Woodlark plate (present-day eastern New Guinea) in the late Oligocene (when most of New Guinea was submerged) and, subsequently (throughout the Miocene), it migrated westwards (to the Maoke and Bird's Head Plates and thereon) and further diversified, in agreement with previous reconstructions.

Keywords: sequence capture, target enrichment, herbariomics, historical biogeography, Papuasia, New Guinea, Araliaceae, *Schefflera*

INTRODUCTION

Situated at the crossroads between Asia and Australia (Lohman et al., 2011), Papuasia has inspired biogeographic research since the time of Wallace (1869). Plate tectonic processes, from volcanism and deformation to ophiolite obduction and island-arc accretion (Baldwin et al., 2012), have created a plethora of terrestrial ecosystems – from mangroves to subalpine grasslands, through tropical forests (Paijmans, 1976; Wikramanayake, 2002; Marshall, 2007) – that support some of the richest diversity on Earth (Williams, 2011). The Papuasian floristic region comprises the main island of New Guinea, the Bismarck Archipelago, and the Solomon Islands (Warburg, 1891; Brummitt, 2001). Its 54% endemic plant taxa (van Welzen et al., 2011) is attributed to its high environmental heterogeneity and isolation (Beehler, 2007; Mutke et al., 2011). Indeed, elevation and terrain ruggedness (i.e., elevational heterogeneity) have been shown to strongly correlate with orchid diversity (Vollering et al., 2016) and even terrestrial-plant genus richness (Hoover et al., 2017), hinting toward orogenetic (i.e., mountain building) processes as catalysts of plant radiation in the region. The spatial distribution of morphological clades in families Sapindaceae (van Welzen et al., 2001) and Ericaceae (Heads, 2003) broadly correspond to geological terranes (i.e., crust fragments sutured to a plate other than that of origin) of various ages leading the authors of both studies to ascribe cladogenesis to vicariance events. Despite these apparent associations, the evolutionary links between past geological events and present-day distributions remain largely unexplored in the region.

With over 600 species, *Schefflera* J.R. Forst & G. Forst s.l. is one of the largest angiosperm genera and the most speciose genus in Araliaceae (Frodin, 1975; Frodin and Govaerts, 2003; Frodin, 2004). In Papuasia, the genus has around 200 estimated species and exhibits a wide environmental tolerance (**Figure 1**) and plasticity of growth forms. *Schefflera* s.l. attains its greatest diversity in Papuasia as trees in montane forests between 1,000 and 2,500 m a.s.l., but other growth forms also include prominent epiphytes or shrubs in lower-montane (650–1,500 m) to mid- and upper-montane forests (1,500–3,200 m; Johns et al., 2007b) or even canopy-emergent trees in sub-alpine ecosystems (3,200–4,200 m; Brass, 1941; van Royen, 1979; Johns et al., 2007a), making *Schefflera* an ideal case study to investigate woody angiosperm diversification in Papuasia. Tectonic models and stratigraphic evidence indicate that New Guinea's mountains attained their present height by rapid uplift from the late Miocene to the early Pliocene (van Ufford and Cloos, 2005; Hall, 2009). While this suggests that Papuasian *Schefflera* rapidly diversified within the last 10–5 Myr, studies that test this hypothesis using a representative sample of Papuasian *Schefflera* have so far been wanting (i.e., nine accessions in Li and Wen, 2014).

Morphologically, most Papuasian *Schefflera* are classified into three largely endemic infrageneric groups: Brassaia, Papuoschefflera, and Pagiophyllae (Plunkett et al., 2005). The first two have overlapping distributions, while the third one is restricted to montane forests and grasslands above 2000 m in the Maoke Mountains (Papua prov., Indonesia). Papuoschefflera is further classified into six provisional morphological groupings which we here name "Bougainvilleanae," "Ischnoacrae," "Versteegiae," "Schumannianae," "Ischyrocephalae," and "Morobeae" (Papuoschefflera one through six, respectively, in Frodin et al., 2010), most of which are geographically restricted within Papuasia. Other morphogroups in the region include Malesian Heptapleurum, Pacific *Plerandra*, and Philippine-Papuasian Cephaloschefflera. *Schefflera elliptica* (Heptapleurum group) is widespread in Southeast Asia, while *Plerandra* and Cephaloschefflera mostly occur around south-eastern New Guinea.

Unraveling the evolutionary history of Papuasian *Schefflera* requires time-calibrated molecular phylogenies both to ascertain the monophyly of these morphologically defined groupings and to shed light on their observed geographic ranges (Wen et al., 2013).

Phylogenies reconstructed from a nuclear region (ITS) and a plastid locus (*trn*L-*trn*F) indicate that *Schefflera* s.l. is polyphyletic, comprising five geographically distinct clades (Plunkett et al., 2005). Using five additional plastid regions, Li and Wen (2014) established the monophyly of nine accessions from western New Guinea and dated their divergence from Heptapleurum to have taken place ∼22 Ma. However, the Papuan clade recovered by Li and Wen (2014) included *S. lucescens* and *S. polybotrya*, which have been recorded from Java, Sumatra, and Borneo but not Papuasia. Frodin et al. (2010) placed *S. lucescens* in the mainly Sundan Paratropia group, together with *S.* "*zollingeriana*" sp. ined., which Plunkett et al. (2005) had resolved as the sister group to Papuasian *Schefflera*. As a result, two competing hypotheses on the origins of Papuasian *Schefflera* exist: either they originated within or outside of Papuasia.

Hyb-Seq (Weitemier et al., 2014) is a high-throughput sequence capture approach that combines target enrichment of, e.g., low-copy nuclear orthologs with genome skimming (additionally yielding high copy number regions, like organellar ones). Whereas previous molecular phylogenies of Papuasian taxa relied almost exclusively on silica-dried samples, Hyb-Seq can be implemented on DNA of varying quality, including that of range-restricted and under-collected species only accessible as herbarium specimens (Hart et al., 2016). In this way, Hyb-Seq has been successfully implemented at the population level and above, even with degraded DNA from centuries-old specimens (Villaverde et al., 2018; Brewer et al., 2019). Still, Hyb-Seq, and other target enrichment techniques, have yet to be widely adopted due to the high cost, prior knowledge needed (e.g., available transcriptomes or genomes), and expertise required (e.g., computational skills) at the initial design and optimization stages of taxon-specific probes (Lemmon and Lemmon, 2013; Dodsworth et al., 2019). To overcome this design and optimization hurdle, the Angiosperms-353 probe set was developed from available transcriptome and genome data (One Thousand Plant Transcriptomes Initiative, 2019) – including twelve Apiales representatives, three of them (*Hedera*, *Hydrocotyle*, and *Polyscias*) in Araliaceae – with universal probes that capture 353 nuclear single-copy loci shared across all angiosperms (Johnson et al., 2018). By using the Angiosperms-353 sequence capture probes, our study is among the first to test the efficacy of this probe set to resolve species-level relationships

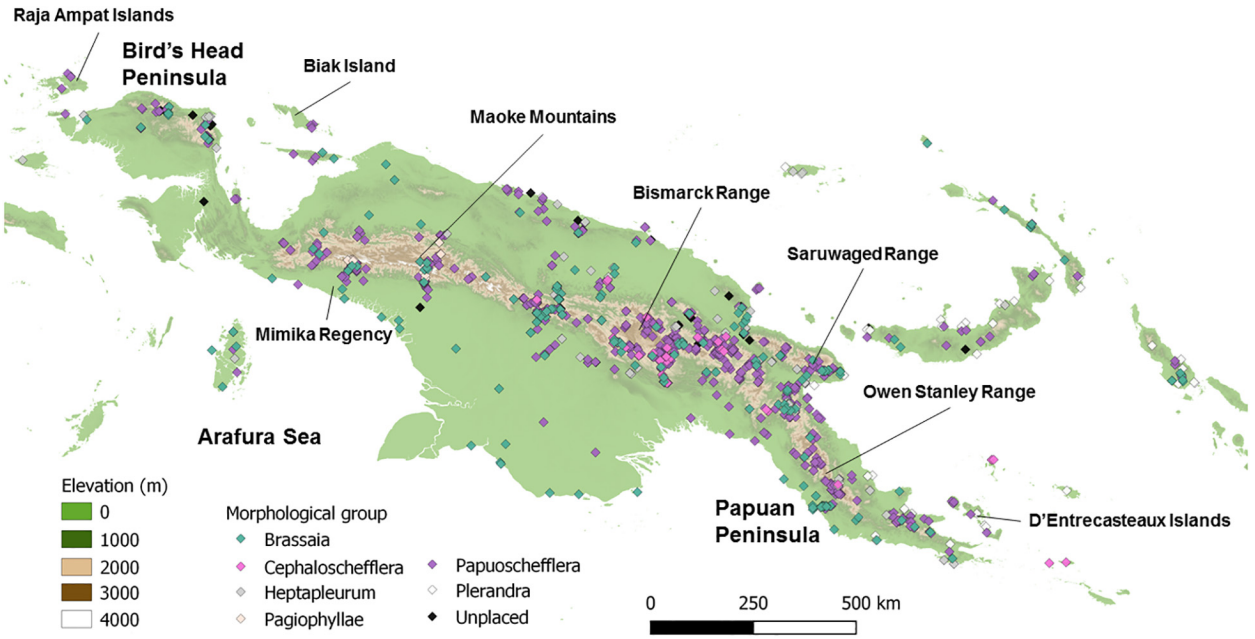

FIGURE 1 | Distribution of *Schefflera* s.l. collections in Papuasia. Over 2,000 collections (colored diamonds) are known from the region, across an elevation gradient from 0 to 4025 m a.s.l. Collections are color-coded according to morphologically defined groups.

from either historical herbarium specimens (Brewer et al., 2019; Larridon et al., 2020; Murphy et al., 2020) or fresh tissue (Van Andel et al., 2019).

Here, we aim to reconstruct the evolutionary history of Papuasian *Schefflera*. To do so, we ask whether its morphologically defined infrageneric groupings are monophyletic, when these lineages diverged, and where (within Papuasia or elsewhere) they diversified.

MATERIALS AND METHODS

Taxon Sampling

We sampled foliar tissue (i.e., lamina, petioles) from 195 herbarium specimens collected between 1850 and 2018 (including four type specimens). These were selected from a georeferenced database of *Schefflera* s.l. compiled by D.G. Frodin, with the addition of four samples from the Royal Botanic Gardens, Kew (RBGK) DNA Bank[1], and four silica-preserved tissue samples from the RBGK Living Collection[2] and the Raja Ampat Islands, West Papua, Indonesia (J. Schrader, GAU Göttingen). Of the 203 sampled specimens, only 74 could be successfully sequenced (due to funding constraints), though these are representative of all Papuasian morphogroups (Frodin et al., 2010) and cover the entire geographic range of the genus in Papuasia (**Figure 1** and **Supplementary Data Sheet S1A**). We selected outgroup taxa from the four primary clades of Asian *Schefflera* (Li and Wen, 2014), plus its closest Araliad relatives: *Heteropanax fragrans* (Roxb.) Seem and *Tetrapanax papyrifer* (Hook.) K.Koch. We also sequenced geographic clades of *Schefflera* that diverged earlier than Asian *Schefflera* (Nicolas and Plunkett, 2014; **Supplementary Data Sheet S1B**) and included genomic sequences of taxa from other major clades in Araliaceae, available through the Plant and Fungal Tree of Life (PAFTOL) Research Programme (Johnson et al., 2018) and the 1000 Plants (1KP) Initiative (Matasci et al., 2014; **Supplementary Data Sheet S1C**).

Laboratory Protocols

DNA Extraction

Samples were washed with 70% ethanol, then kept at −80°C for 12 h (to facilitate cell wall breakage) and milled in an MM400 (Retsch Inc.) grinder. Genomic DNA was extracted using a modified-CTAB protocol (Doyle and Doyle, 1987), further adjusted to improve yield from herbarium samples (**Supplementary Data Sheet S2A**). Key changes include incubating samples in CTAB at 65°C for 12 h (to optimize DNA isolation) and in isopropanol at −20°C for 48 h (to improve precipitation of fragmented DNA). Precipitated DNA pellets were washed twice with 70% ethanol and resuspended in Milli-Q ultrapure (Type 1) water (Merck KGaA). We measured relative DNA concentration (ng/μL) with the QuantiFluor® dsDNA System (Promega Corp.). Samples were purified using Agencourt AMPure XP beads (Beckman Coulter Life Sciences) (**Supplementary Data Sheet S2B**). Extractions from pre-1970 collections or with concentrations <10 ng/μL were treated as having highly fragmented DNA and were cleaned using AMPure beads and isopropanol following Lee (2014) to reduce loss of

[1]https://www.kew.org/data/dnaBank/
[2]http://www.kew.org/data/lcd.html

small (<300 bp) DNA fragments (Särkinen et al., 2012). Where material was available, we repeated extractions to obtain at least 200 ng of DNA from each sample. DNA fragment size distribution was determined by gel electrophoresis (**Supplementary Data Sheet S3A**). Extractions with fragments predominantly above the target insert size (≥ 500 bp) were sonicated with an ME220 Focused-ultrasonicator™ (Covaris Inc.).

Genomic Library Preparation

We prepared genomic libraries using the NEBNext® Ultra™ II DNA Library Prep Kit and Multiplex Oligos (Dual Index Primers, sets 1 and 2) for Illumina® (New England Biolabs) at half-volumes to reduce per sample costs. Target insert size was 350 bp and size selection was not required where DNA template was highly degraded (<300 bp). Size-selected libraries were amplified with 13 PCR cycles and all others with 14 cycles. Libraries were re-amplified, where required, using KAPA HiFi HotStart ReadyMix (Roche) with i5 and i7 forward and reverse primers (as described in Meyer and Kircher, 2010) to obtain at least 75 ng/library.

Hybridization and Sequencing

Libraries were multiplexed (11–12 per pool) for hybridization. Equimolar library pools were made homogenizing phylogenetic distances (if known) and avoiding combinations of libraries originating from different quality DNA to reduce uneven bait competition within hybridization pools. We considered the following criteria: (i) whether re-amplification was required, (ii) whether sonication was required, (iii) whether size selection was required, and (iv) whether there was sufficient library template (**Supplementary Table S3B**). Outgroup taxa were pooled separately from Papuasian taxa where possible to even out occupation of bait binding sites. Each pool contained 500–1,000 ng DNA in total.

Pools were enriched using the Angiosperms-353 myBaits® Expert Panel target capture kit (Arbor Biosciences). They were hybridized at 65°C for 24 h, then amplified with i5 and i7 forward and reverse primers (Meyer and Kircher, 2010) for 14–18 PCR cycles to obtain pools at least 1 nM. Each pool was quality-controlled with a TapeStation 4200 (Agilent Technologies) (**Supplementary Data Sheet S3C**). Due to funding constraints, we only sequenced the eight library pools with the highest quality, which covered the broadest diversity range and consisted of 90 accessions in total (though only 74 passed quality filters). These were denatured and diluted following manufacturer's specifications (Illumina® protocol # 15039740) and loaded at 16 pM for sequencing in an Illumina® MiSeq using two v2 (300-cycles) reagent kits (Illumina®, Inc.) at the Jodrell Laboratory (Royal Botanic Gardens, Kew, Richmond, United Kingdom).

Bioinformatic Analyses

Sequence Rescue and Alignment

Sequences were trimmed using Trimmomatic 0.38 (Bolger et al., 2014), employing "palindrome mode" adapter removal and Maximum Information quality filter settings to favor longer reads (ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:2:TRUE MAXINFO:40:0.2 LEADING:3 TRAILING:3 MINLEN:36). We examined sequence quality with FastQC 0.11.7 (Andrews, 2018) before and after trimming to ensure complete adapter removal and to identify surviving artifacts that could affect downstream analyses.

HybPiper 1.3.1 (Johnson et al., 2016) was used to retrieve target sequences of nuclear genes (exonerate script) and flanking off-target regions (intronerate.py script, which reruns exonerate but, instead of removing flanking regions, it keeps them), with the BWA mapper (Li, 2013) and the SPAdes assembler (Bankevich et al., 2012) (–bwa -cov-cutoff = 3). We investigated polymorphic sequences for paralogy using exploratory trees built from MAFFT 7.215 (Katoh and Standley, 2013) (–auto) alignments in FastTree 2 (Price et al., 2010) (-nt -gtr). We discarded any sequence that may have resulted from gene duplication and retained the most common allele when sequences were homologous (Kates et al., 2018). We used our own custom script, *max_overlap*[3] (**Supplementary Data Sheet S4A**), to calculate a coverage score for each sequence that is proportional to representativeness (proportion of accessions/loci with sequences), data matrix completeness (percent sequence recovered against overall length of target), and evenness of distribution (adapted from Pielou, 1966) across accessions per locus and so, to reduce noise in our data matrixes by filtering out underrepresented, incomplete, and unevenly distributed sequences.

Internal Transcribed Spacer (ITS1-5.8S-ITS2) nuclear ribosomal DNA sequences (hereafter ITS) were also retrieved with HybPiper using a target file we made from aligned ITS sequences – from Li and Wen (2014) and Plunkett et al. (2004, 2005) – deposited in GenBank (Benson et al., 2018). Off-target sequences corresponding to 142 plastid loci (genes and intergenic spacers) were similarly rescued with HybPiper using a plastid-target file we generated from the complete plastid genomes of *S. heptaphylla* (L.) Frodin (Zong et al., 2016: KT748629), *Aralia elata* (Miq.) Seem. (Kim and Yang, 2016: KT153023), *S. delavayi* (Franch.) Harms, and *Metapanax delavayi* (Franch.) J.Wen & Frodin (Li et al., 2013: KC456166, KC456165).

All sequences were aligned with UPP (Nguyen N.-P.D. et al., 2015) to produce accurate alignments from fragmentary datasets (i.e., historical herbarium DNA template), using only those with >95% of the longest available sequence length for the backbone dataset. UPP uses hidden Markov models (HMM) for multiple sequence alignment (MSA) and it relies on PASTA 1.8.4 (Mirarab et al., 2015) (a divide-and-conquer MSA method) to generate initial backbone alignments. In turn, PASTA relies on FastTree 2 (Price et al., 2010) for tree estimation, MAFFT 7.215 (Katoh and Standley, 2013) for alignment, and OPAL 2.1.3 (Wheeler and Kececioglu, 2007) for merging. We trimmed alignments using our own custom script, *optrimAl*[4] (**Supplementary Data Sheet S4B**), which optimizes the gap threshold value in trimAl 1.2 (Capella-Gutiérrez et al., 2009), to obtain the highest proportion of parsimony-informative characters (P_{PIC}) while retaining adequate sequence length (Shen et al., 2016). Alignments where

[3]https://github.com/keblat/bioinfo-utils/tree/master/docs/advice/scripts/max_overlap.R

[4]https://github.com/keblat/bioinfo-utils/tree/master/docs/advice/scripts/optrimAl.txt

trimming resulted in data loss exceeding 30% were interpreted to contain low ratios of phylogenetic signal to noise and were discarded. Alignment statistics were calculated in AMAS 0.98 (Borowiec, 2016) using the summary function. Sequence capture statistics were calculated using R 3.5.3 (R Core Team, 2019).

Gene and Species Tree Inference

Gene trees were inferred with IQ-TREE 1.6.10 (Nguyen L.-T. et al., 2015) after selecting substitution models with ModelFinder (-m TEST) (Kalyaanamoorthy et al., 2017). Outlier branches that increased the diameter of each gene tree by more than 20% were identified using TreeShrink 1.3.1 (Mai and Mirarab, 2018) with centroid re-rooting (-b 20 -c) and removed. Each locus was then realigned without the outlier sequences. We estimated bipartition support with 1000 UFBoot2 (Hoang et al., 2018) bootstrap replicates using IQ-TREE (-bb 1000) and contracted branches with support values below 10% (Mirarab, 2019) using Newick Utilities 1.6 (nw_ed "i & b ≤ 10") (Junier and Zdobnov, 2010).

Pseudo-coalescent species trees were inferred using ASTRAL III v5.6.3 (Zhang et al., 2018). Species trees were inferred separately for the nuclear and chloroplast genomes as they represent different evolutionary pathways (Ravi et al., 2008). We used the local posterior probabilities (PP_{local}) calculated in ASTRAL to estimate quartet support for the recovered topology at each node. Conflict, concordance, and phylogenetic signal were assessed with phyparts (Smith et al., 2015) and displayed with the PhyPartsPieCharts script[5], which depicts the number of gene trees that support, oppose or provide no information with respect to the dominant species tree topology. Unresolved polytomies in the final species tree were tested in ASTRAL (-t 10) to determine if they are due to insufficient data or could possibly reflect a true polytomy (Sayyari and Mirarab, 2018).

To verify placement of our sampled Papuasian and outgroup accessions within family Araliaceae, we inferred the ITS gene tree for these together with sequences of other Araliaceae accessions from Li and Wen (2014) and Plunkett et al. (2004, 2005). We used two accessions from Myodocarpaceae (*Delarbrea paradoxa* Vieill. and *Myodocarpus fraxinifolius* Brongn. & Gris.) as the outgroup. The gene tree was estimated in MrBayes 3.2.6 (Ronquist et al., 2012) under a GTR + Γ substitution model (with settings: nchains = 4, nruns = 4, and MCMC ngen = 100M).

Divergence Time Estimation and Ancestral Area Reconstruction

To limit variation in substitution rates and minimize overestimation of recent divergence times (Ho et al., 2005), we selected nuclear genes that (i) were at least 10% concordant with the species tree (bipartition support > 0.1), (ii) likely evolved according to a strict molecular clock (we set root-tip variation to <0.024), (iii) contained the most information (tree length > 0.1), and (iv) represented the most accessions (at least 25) with SortaDate (Smith et al., 2018). To reconstruct the biogeographic history of Papuasian *Schefflera* (Crisp et al., 2011), we included in the data matrix ITS sequences from other Asian *Schefflera* clades sampled by Li and Wen (2014). All other loci for these Asian accessions were coded as missing data. This resulted in a 51-taxa data matrix, partitioned by locus (we used independent substitution models for each locus, as selected by ModelFinder), consisting of only Asian and Papuasian *Schefflera* sequences. This data matrix comprised 31,496 bp across 10 nuclear exonic regions (6,469 bp), 15 nuclear flanking regions (24,177 bp), and nuclear rDNA ITS. The data matrix had 11.6% parsimony-informative sites and, as 16 taxa (32% of the total sampling) are represented only by ITS sequences, 55.5% missing data.

Divergence times were estimated in BEAST 1.10.4 (Suchard et al., 2018) on the CIPRES Science Gateway v3.3 online platform (Miller et al., 2011). Since known Araliaceae fossils lay outside our focus group, three secondary time constraints – drawn from Li and Wen (2014) – were imposed on: (i) the root node (normal prior distribution with mean = 42.0 and st.dev. = 8.0); (ii) the Heptapleurum crown node (normal, mean = 36.0, st.dev. = 7.0); and (iii) the Papuasian *Schefflera* crown node (normal, mean = 22.0, st.dev. = 5.0). To reduce search-space and avoid miss-rooting problems with BEAST analyses, we enforced monophyly on the Agalma, Brassaia, *S. elliptica* alliance, Heptapleurum, Heptaphylla + Hypoleuca, Ischyrocephalae, and Papuoschefflera s.s. clades (fully supported in our species tree), as well as the root node, to prevent inverted ingroup-outgroup topologies (for further details see Springer et al., 2018).

Concurrently, ancestral areas were reconstructed using a fully probabilistic approach – first described by Sanmartín et al. (2008) and implemented in BEAST by Lemey et al. (2009) – that infers diffusion processes among discrete locations in timed evolutionary histories under Bayesian stochastic search variable selection (BSSVS). Continuous-time Markov chains (CTMC) are used to model instantaneous geographic locations of any given sequence, together with the transition and migration rates between these locations. Since this Bayesian CTMC phylogeographic model assumes ancestral ranges are limited to single regions, it requires discretizing the entire distribution of any given taxon (i.e., a given taxon can be represented by multiple accessions), while tolerating incomplete sampling (Drummond et al., 2012). Moreover, unlike other approaches (e.g., dispersal-vicariance parsimony or dispersal-extinction cladogenesis), it can be implemented in scenarios where the number of areas is large (>10 areas), allowing for fine-scale area explorations (e.g., Mairal et al., 2015, 2017). Thus, we included a partition with collection localities – coded according to tectonic plate boundaries (Bird, 2003) – in our BEAST input file (generated in BEAUTi 1.10.4; Suchard et al., 2018) and we tested six models, comprising all possible combinations of three clock priors – strict, random local, and uncorrelated relaxed lognormal – and two species-tree priors robust to incomplete sampling (our case) – Yule process (Yule, 1925) and birth-death incomplete sampling (Stadler, 2009). We selected the best-supported model by estimating marginal likelihoods (MLEs, path steps = 100, chain length = 1M), using path sampling (PS) and stepping-stone sampling (SS) (Baele et al., 2013), from runs that converged after 100M iterations. BEAST log files were loaded into Tracer 1.7.1 (Rambaut et al., 2018) and visually inspected to check that the chains had converged, and that mixing and Effective

[5]https://github.com/mossmatters/MJPythonNotebooks

Sample Sizes (ESS > 200) were adequate for all parameters (after 100M iterations). After discarding burn-in iterations, trees were annotated and posterior probabilities (*PP*) summarized in TreeAnnotator 1.10.4 (Suchard et al., 2018) on the tree in the posterior sample with the maximum sum of the posterior clade probabilities (MCC tree), rescaling to reflect median node heights for clades contained in said tree. The resulting MCC tree was visualized in FigTree 1.4.4 (Rambaut, 2018).

RESULTS

Sequence Capture Success and Bioinformatic Analyses

Of the 90 accessions sequenced, only 74 had sufficient reads after quality filtering for target retrieval with HybPiper (median = 737,691 reads/sample; **Supplementary Table S5A**). We find that specimen age had no significant effect on the number of reads and that read yield did not differ between herbarium specimens and other material (i.e., Kew DNA bank and silica-dried samples; **Figure 2**). For nuclear loci in the Angiosperms-353 enrichment panel, on average 11.5% reads/sample were on-target. Capture success (defined as the proportion of total reference sequence recovered) varied widely (range = 0.1 – 64.6%) across samples for herbarium material (median = 28% reads/sample) and was weakly correlated with specimen age (F = 7.01, DF = 66, $p < 0.01$, R^2 = 0.08). DNA bank and silica samples yielded higher capture success (t = 14.1, DF = 71.6, $p < 0.001$). For off-target plastid regions (including both coding loci and intergenic spacers), on average 16.9% of reads/sample mapped to our "plastid-target" custom file, with 58% median "capture success" for herbarium material. Plastid "capture success" was nearly complete for 16 samples (**Supplementary Table S5B**) and was weakly correlated with specimen age (F = 6.34, DF = 66, $p < 0.01$, R^2 = 0.07). In total, we obtained sequences for 352 coding (on-target) and 349 flanking (off-target) regions from the nucleus, and 73 coding and 64 intergenic spacers from the chloroplast (off-target), as well as the nuclear rDNA ITS region (recovered with "ITS-target" custom file). Sixty potential paralogs were detected based on gene tree topology (**Supplementary Table S5C**), of which 23 nuclear and two plastid genes were probable duplications and excluded from downstream analyses.

The proportion of parsimony-informative characters (P_{PIC}) was highly variable between multiple sequence alignments (MSA) for nuclear coding loci and flanking regions (**Table 1**). P_{PIC} was generally low for off-target plastid genes and intergenic spacers. Non-coding off-target regions (nuclear flanking regions and plastid spacers) had higher P_{PIC} than their respective coding counterparts (nuclear targets and off-target plastid loci). Missing data accounted for 27.1% of nuclear targets, 42.3% of nuclear flanking regions (off-target), 7.2% of plastid loci, and 8.0% of plastid spacers. No phylogenetic bias was observed in the distribution of missing data. We included only nuclear regions (both coding and flanking) for phylogenetic inference as the low levels of informativeness in the plastid sequences were more likely to lead to gene tree estimation errors (Meiklejohn et al., 2016). Only 39 accessions had a coverage score of at least 0.5 for nuclear sequence capture; this coverage score is proportional to representativeness x completeness x evenness (see Methods above). These 39 accessions, together with three sequences from 1KP (One Thousand Plant Transcriptomes Initiative, 2019) and one sequence from PAFTOL (Johnson et al., 2018), were used for downstream phylogenomic analyses (pseudo-coalescent framework). The final data set included 33 Papuasian *Schefflera* accessions and 12 outgroup accessions. After trimming, the data set to be used in pseudo-coalescent analyses comprised 354,057 bases (**Supplementary Table S5D**) across 141 nuclear coding regions (93,325 bp) and 163 nuclear flanking regions (260,732 bp).

Phylogenetic Relationships in *Schefflera*

The monophyly of most of the currently accepted genera in Araliaceae was strongly supported in the Bayesian ITS tree (**Figure 3** left), save *Polyscias* (sensu Lowry and Plunkett, 2010), which had relatively low support. *Panax* was nested within *Aralia* and *Chengiopanax* was nested in *Gamblea*. As expected, *Schefflera* was highly polyphyletic, following the geographical clades of Plunkett et al. (2005). *Schefflera*'s "Asian Palmate clade" (Plunkett et al., 2005) was strongly supported (with Neotropical *Schefflera* nested within it) and had *Tetrapanax papyrifer* and *Heteropanax fragrans* gradually leading to the "Asian *Schefflera* clade" with maximum support. Within this latter "Asian *Schefflera* clade," all major clades (sensu Li and Wen, 2014) were also strongly supported including Heptapleurum, which comprises the *Schefflera elliptica* alliance, the Philippine *Schefflera*, and the Papuasian clade, which has *Schefflera tristis* as sister lineage. The pseudo-coalescent species tree also recovered monophyletic Papuasian *Schefflera* nested within Heptapleurum, which was itself nested in the "Asian *Schefflera* clade" (**Figure 4** left) with maximum quartet support.

Within Papuasian *Schefflera* (**Figure 3** right, **Figure 4** right, and **Figure 5**), the monophyly of the Brassaia morphogroup was strongly supported, regardless of the inference approach taken. However, phylogenetic relationships within Brassaia were poorly resolved. The Papuoschefflera morphogroup was paraphyletic with regard to Brassaia in the ITS tree, it was supported as sister to Brassaia in the pseudo-coalescent tree and, although the latter topology was also retrieved in the chronogram, it had low support. All three trees disagreed on the placement of Cephaloschefflera (*S. eriocephala*) but found Pagiophyllae (*S. "frigidariorum"* sp. ined.) to be nested within Papuoschefflera. Hereafter, we refer to the taxa that are not part of the Brassaia clade as Papuoschefflera s.l. and restrict the circumscription of Papuoschefflera s.s. to Pagiophyllae plus Bougainvilleanae, Schumannianae, and Versteegiae. These morphogroups were reconstructed in a highly supported clade in all trees and are generally distributed across the western half of Papuasia (**Figure 5** bottom left). Morphogroup Ischyrocephalae had maximum support both in the ITS tree and the chronogram but had *S.* "*goodenoughiana*" sp. ined. (Oreopolae, D.G. Frodin pers. comm.) nested in the pseudo-coalescent tree, which resulted in a paraphyletic

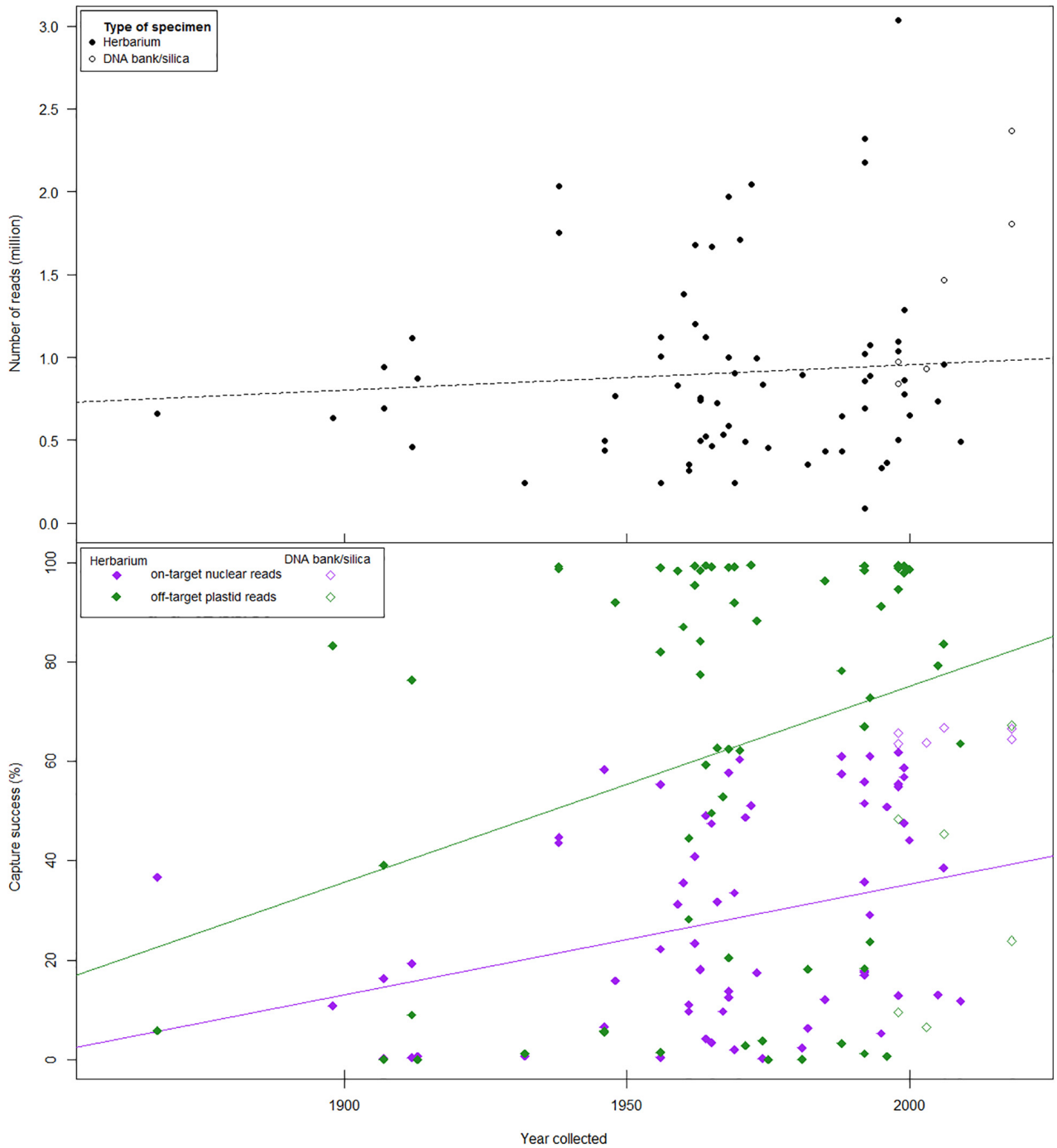

FIGURE 2 | Reads mapped and capture success with respect to specimen age and template type (DNA bank/silica vs. herbarium). **(Top)** The number of reads obtained did not differ between herbarium specimens (filled circles) and DNA bank/silica samples (empty circles). There was no correlation between the number of reads and specimen age (dotted black line). **(Bottom)** Capture success (proportion of total reference sequence recovered) of nuclear genes (empty purple diamonds) was significantly higher in DNA bank/silica samples than in herbarium specimens (filled purple diamonds). Capture success in herbarium samples was weakly correlated with specimen age for both nuclear genes (purple line) and plastid loci (green line).

Ischyrocephalae (and a polyphyletic Oreopolae). Monophyly of morphogroups Ischnoacrae and Morobeae could not be tested as sequenced representative samples did not yield sufficient reads on target.

At the species level, all morphological taxa sampled (except for *S. actinophylla*) were reconstructed as monophyletic, though support was variable. *S.* "*aeruginosa*" sp. ined., *S.* "*frigidariorum*" sp. ined., *S. pilematophora*, and *S. pueckleri* had high support in

TABLE 1 | Alignment statistics across retrieved regions.

Genomic compartment	Region	Alignment length		P_{PIC}*		Missing data	
		Mean (bp)	*SD*	Mean (%)	*SD*	Mean (%)	*SD*
Nuclear	On-target coding	613	368	17.0	10.4	27.1	11.8
	Off-target flanking	886	484	28.2	14.3	42.3	10.5
Plastid	Off-target coding	873	830	1.7	1.7	7.2	5.2
	Off-target non-coding	534	531	2.7	2.2	8.0	5.5
All		696	506	20.3	13.6	47.6	15.6

*P_{PIC}, *Proportion of Parsimony-informative characters.*

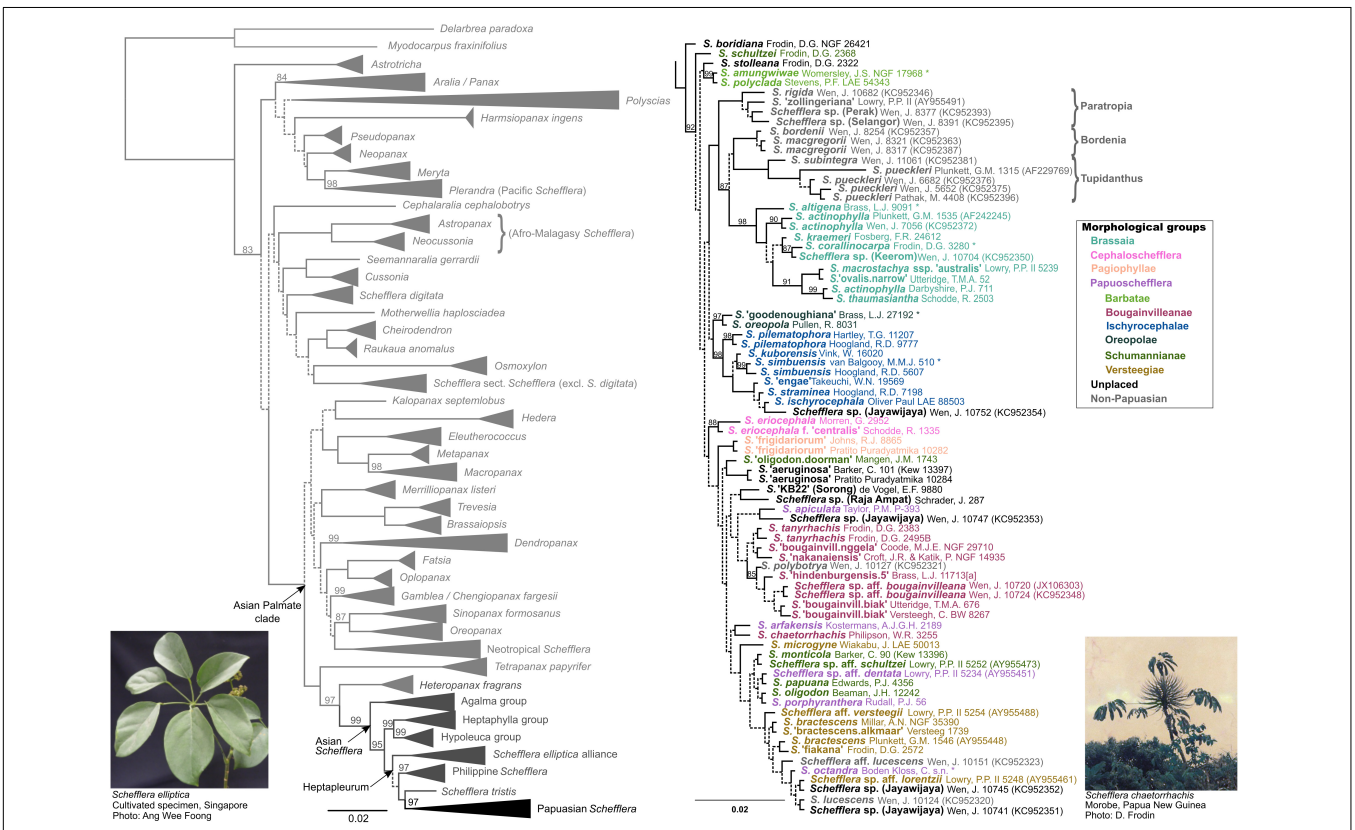

FIGURE 3 | Bayesian ITS gene tree (MrBayes). Dashed lines represent low confidence (PP_{ITS} < 0.8). Solid lines have maximum support (PP_{ITS} = 1) unless otherwise stated (percentages by branches). **(Left)** Araliaceae genera. All generic clades have been collapsed in this tree, including Papuasian *Schefflera*. **(Right)** Papuasian *Schefflera*. Accession labels are color-coded according to infrageneric morphogroups as in Frodin et al. (2010).

all trees. Support for *S. simbuensis* was high in the chronogram and the ITS tree but low in the pseudo-coalescent tree. Similarly, *S. eriocephala* was well supported in the chronogram but not so in the ITS tree. *Schefflera actinophylla* had two non-Papuasian accessions group together to form a well-supported clade in the ITS tree and a third Papuasian accession highly supported as sister to also Papuasian *S. thaumasiantha*.

Divergence Time and Ancestral Area Co-estimation

Speciation under a non-coalescent Yule process and an uncorrelated relaxed molecular clock (lognormal distribution) was selected as the best-fit model (**Table 2**). Heptapleurum, with an early Oligocene crown age of ∼33.4 Ma, was inferred to have originated in the Sunda plate (**Figure 5** right). Heptapleurum transitioned from Sunda into the Woodlark plate sometime in the late Oligocene, between ∼29 and 26.3 Ma, giving rise to the Papuasian *Schefflera* clade.

Within the Papuasian clade, Brassaia was reconstructed as having originated in the Sahul shelf, having arrived from the Woodlark plate in the early Miocene, between ∼23.8 and 18.1 Ma. Paratropia, Bordenia, and Tupidanthus (*S. rigida, S. macgregorii,* and *S. subintegra*, respectively) are here monophyletic and sister to Brassaia, though with low support. These three morphogroups were reconstructed to have

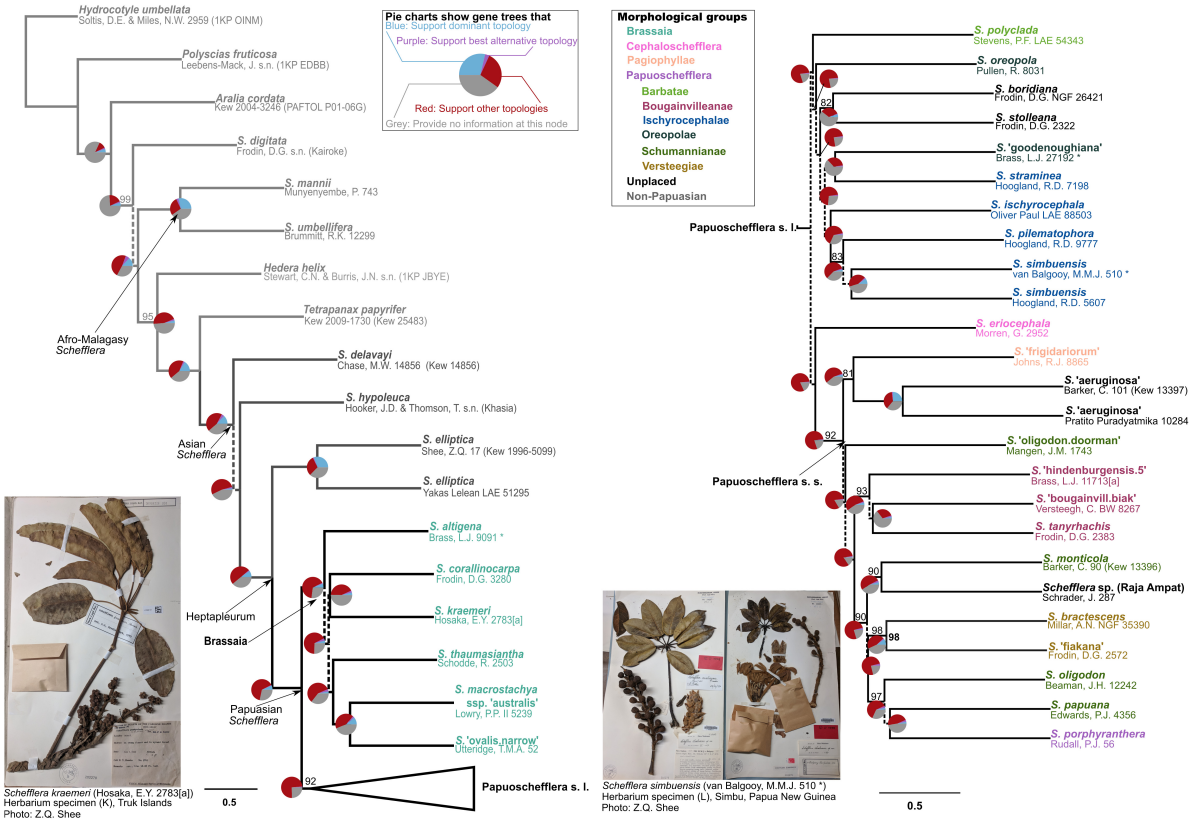

FIGURE 4 | Pseudo-coalescent species tree of Papuasian *Schefflera* (ASTRAL). Pie charts show gene tree concordance with coalescent species tree. Dashed lines represent low confidence (PP_{local} < 0.8). Solid lines have maximum support (PP_{local} = 1) unless otherwise stated (percentages by branches). **(Left)** Outgroup taxa and Papuasian subgeneric groupings. Papuoschefflera has been collapsed in this tree. **(Right)** Papuoschefflera infrageneric groupings. Accessions are color-coded according to infrageneric morphogroups as in Frodin et al. (2010).

transitioned back to Sunda, from Woodlark, in the Early Miocene (between ∼23.8 and 16 Ma) and, from there, onward to the Philippines sometime between ∼13.4 Ma and the present. An additional dispersal event to the Philippines, this time from Sunda, took place sometime between ∼33.5 Ma and the present and resulted in the Philippine *Schefflera* clade (*S. blancoi*). Morphogroup Ischyrocephalae was inferred as sister to *S. boridiana* plus *S. stolleana*. Crown age for Ischyrocephalae is ∼17.6 Ma and it reached the South Bismarck plate, from the Woodlark plate, twice between ∼11.6 Ma and the present. Morphogroup Oreopolae (*S. oreopola* and *S.* "*goodenoughiana*" sp. ined.) was reconstructed in a clade with maximum support and as sister to Barbatae (*S. polyclada*). This well-supported Oreopolae + Barbatae clade transitioned into the Sahul shelf from Woodlark in the Middle Miocene, between ∼17.3 and 12.6 Ma, returning to Woodlark between ∼7.43 Ma and the present. Additionally, the Oreopolae + Barbatae clade converged with an also well-supported Cephaloschefflera clade (*S. eriocephala* and *S. eriocephala* f. *centralis*) ∼17.6 Ma to form an Eastern clade. Papuoschefflera s.s. moved from the Woodlark plate to the Maoke plate in the Late Miocene, between ∼23.1 Ma and 20.0 Ma. From there, it colonized the Bird's Head plate multiple times, expanded into the Sahul shelf and re-entered the Woodlark plate.

DISCUSSION

Efficacy of Universal Probes

To date, other than in *Schefflera*, the Angiosperms-353 enrichment panel (Johnson et al., 2018) has only been tested at the species level in sedges (*Cyperus*; Larridon et al., 2020) and Old World pitcher plants (*Nepenthes;* Murphy et al., 2020) and at the population level (SNPs) in rice (*Oryza*; Van Andel et al., 2019). When comparing data matrices containing on-target loci only, both Larridon et al. (2020) and Murphy et al. (2020) had a higher capture success (**Table 3**); an expected result since they worked with a greater proportion of silica-dried tissue (Brewer et al., 2019). However, their mean P_{PIC} was lower than ours (**Table 3**), probably as a result of our filtering approach, which combined two custom scripts – *max_overlap* (representativeness x completeness x evenness) and *optrimAl* (per locus gap threshold optimization) – to increase signal while reducing missing data in our data matrixes. The relatively high informativeness of our final data set suggests that this adaptive trimming may help strike a balance between retaining sequence length and improving phylogenetic informativeness (Hartmann and Vision, 2008).

The low specificity of the Angiosperms-353 baits (probes are <30% divergent; Johnson et al., 2018) would explain why

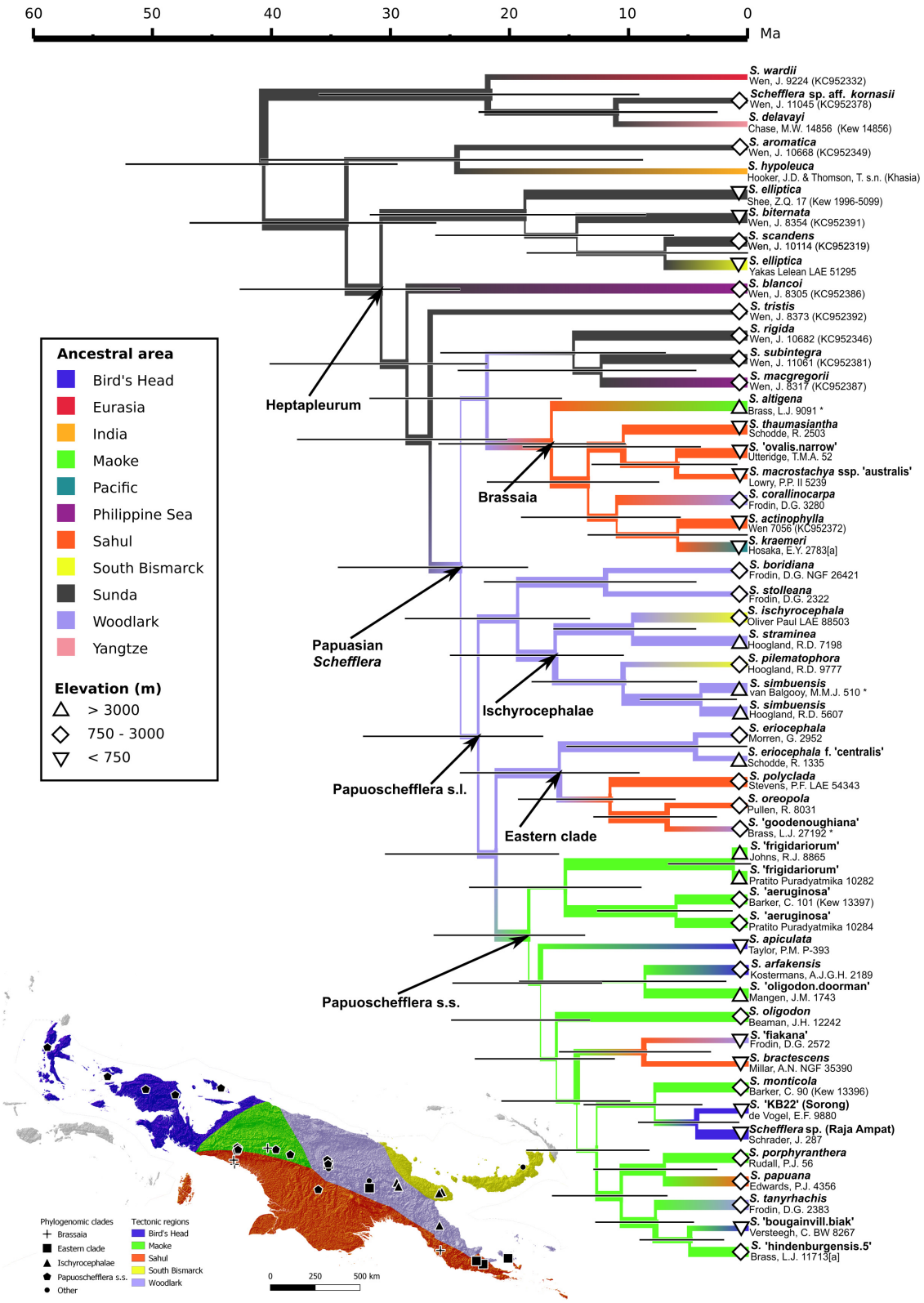

FIGURE 5 | (Bottom left) Distribution of *Schefflera* clades across Papuasia. Accession symbols are coded according to clade. Areas are color-coded according to tectonic plates. The background map is a hillshade of the Digital Elevation Model. **(Right)** Bayesian ancestral area reconstruction for Papuasian *Schefflera* (BEAST). Branches are color-coded according to reconstructed ancestral areas. Symbols next to accession labels indicate collection elevation.

TABLE 2 | Model selection for divergence time estimation and ancestral area reconstruction in BEAST 1.10.4 (Suchard et al., 2018): marginal likelihood estimates (MLEs) for six tree and clock model comparisons.

Tree model	Clock model	Log marginal likelihood	
		Path sampling	Stepping-stone sampling
Yule process	Strict	−127,161.38	−127,159.31
	Random local*	N/A	N/A
	Uncorrelated relaxed lognormal	*−126,125.37*	*−126,125.16*
Birth-death incomplete sampling	Strict	−127,149.93	−127,149.49
	Random local*	N/A	N/A
	Uncorrelated relaxed lognormal	−126,130.93	−126,130.65

**Random local clock models exhibited inadequate chain mixing and did not converge after 100M iterations. Italics indicate best-supported model and associated log MLEs.*

Larridon et al. (2020) and Murphy et al. (2020), and this study retrieve a lower mean percent of on-target reads per sample than other herbariomic studies (Hart et al., 2016; Vatanparast et al., 2018; Villaverde et al., 2018) relying on taxon-specific probes (**Table 3**). Mean capture success of target loci from herbarium specimens in *Schefflera* was comparable to that obtained by Villaverde et al. (2018), which used custom probes, designed for genus *Euphorbia*, rather than universal ones. Increased sequencing depth has been shown to correlate with capture success (Johnson et al., 2018) which, combined with kit specificity and sample age, may explain why mean capture success is higher (than ours) in two legume studies (>80%) using different family-specific bait kits (Hart et al., 2016; Vatanparast et al., 2018). Like Villaverde et al. (2018), we also found that specimen age affected capture success (**Figure 2**), probably due to accumulated DNA damage and its effect on genomic library preparation (Der Sarkissian et al., 2015). The large variance in capture success of post-1940s specimens could be explained in terms of variability in collection, preservation, and storage techniques (Brewer et al., 2019; Forrest et al., 2019).

Other recent studies have also demonstrated the efficacy of taxon-specific probe sets in resolving species-level relationships from herbarium DNA (Finch et al., 2019; Soto Gomez et al., 2019; White et al., 2019). Whereas Kadlec et al. (2017) argued that high sequence variability across angiosperm orders precluded the usefulness of universal probes in resolving species-level relationships, Chau et al. (2018) found that general purpose probes can be as effective as taxon-specific ones. While we do not compare these alternative probes, the results from our study, Larridon et al. (2020) and Murphy et al. (2020) suggest that an appropriately designed universal probe set can capture adequate phylogenomic information to resolve relationships at the species-level and even at the population-level (Van Andel et al., 2019). Liu et al. (2019) showed experimentally that when probes are <30% divergent from regions targeted, enrichment worked adequately (see Figure 6 and Supplementary Figure S6 in Liu et al., 2019). Johnson et al. (2018) took this threshold into account when designing the Angiosperms-353 probe set and included sufficient probes to account for the diversity the panel encompasses (i.e., all angiosperms). Thus, if universal probe sets can indeed be as informative at shallower phylogenetic levels as lineage-specific ones, this would considerably reduce the cost and effort associated with designing and optimizing taxon-specific probes for phylogenomic studies (McKain et al., 2018; Dodsworth et al., 2019).

Phylogenomic Support for Morphological Groupings

The paraphyly of the Papuasian accessions in our study corroborates the previous results of Li and Wen (2014). The Papuasian clade we recovered included accessions from three non-Papuasian lineages: Sundan Paratropia, Philippine Bordenia and mainly Indochinese Tupidanthus (**Figure 3** right). Given our divergence time estimates for the Papuasian clade (**Figure 5** right) and considering the timing of the Sunda-Sahul floristic exchange between ∼34 and 12 Ma (Crayn et al., 2015), Asian *Schefflera* appears to have crossed Wallacea – the floristic province within the Malesia biogeographic region connecting the Sunda and Sahul shelves – at least twice before both these shelves finally merged, supporting the observation made by van Welzen et al. (2011) that "there is no sharp E-W boundary for plant distributions in Malesia" (though see Bacon et al., 2013 for a counter-example in palms).

All our topologies support the current circumscription of Brassaia proposed by Frodin et al. (2010). Brassaia and Papuoschefflera s.l. primarily differ in floral morphology (**Table 4**). An earlier treatment (Harms, 1921) classified Papuasian *Schefflera* into two sections: (i) Cephaloschefflera, with flowers arranged in heads and (ii) Euschefflera, with flowers in umbellules. Our molecular analysis supports the proposal of Frodin (1975) that this character is plesiomorphic in Cephaloschefflera *sensu* Harms (1921). Brassaia was historically treated as a separate genus (Bentham, 1867) until it was incorporated into sect. Cephaloschefflera (Harms, 1921). Later, the four conspicuous floral bracts present in the clade were proposed as an apomorphic character (Frodin, 1975), an observation which appears to be validated by all our topologies (**Figures 3**–**5**).

Within Brassaia, phylogenetic relationships better match geography than morphology. This is exemplified by the strongly supported clade comprising *S. macrostachya* ssp. "*australis*" ssp. ined. and *S.* "*ovalis.narrow*" sp. ined. which was recovered in all three trees (**Figures 3**–**5**). Members of the *S.* "*ovalis*"

TABLE 3 | Comparison of nuclear exon target capture and alignment statistics* across comparable data matrixes (on-target coding only) in herbariomic studies.

Study	Probe set	Target loci	Sample type (Herb/Other)	Collection year	Mean reads per sample	Reads on target (%)	Capture success (%)	Alignment length	Mean P_{PIC} (%)	Missing data (%)
Hart et al., 2016	Leguminosae (specific)	214	11/2	1835–2009	1,241,592	79.9	89	229,995	6.7	22.4
Vatanparast et al., 2018	Leguminosae (specific)	507	12/13	1985–2014	6,356,207	32	81.9	737,309	18.3	19.7
Villaverde et al., 2018	*Euphorbia* (specific)	431	56/88	1891–2014	1,036,822	48.6	26.8	486,878	8	23.3
Soto Gomez et al., 2019	Dioscoreaceae (specific)	260	22/3	1994–2007	922,847	31.6	93.8	276,920	24	5.2
Larridon et al., 2020	Angiosperms (universal)	353	30/8	1948–2017	5,155,095	8.5	33.2	233,429	9.9	40.4
Murphy et al., 2020	Angiosperms (universal)	353	31/194	1835–2019	1,633,509	5.6	59.3	160,320	13.6	9.6
Shee et al., this study	Angiosperms (universal)	353	68/6	1850–2018	946,130	11.5	28.2	194,215	17	27.1

*Calculated from published results, including **Supplementary Material**.

alliance have never been formally described, although their affinity with *S. macrostachya* in leaf venation has been noted (Frodin, 1975). Since the two collections were made within 30 km of each other along the Aikwa River, in Mimika Regency, they may well represent variation within a single species. The same seems to be happening with regards to the polyphyly of *S. actinophylla* in the ITS tree (**Figure 3** right). Our New Guinean *S. actinophylla* accession and *S. thaumasiantha* were from the same locality and formed a well-supported clade. The other clade consisted of a Queensland collection and a cultivated plant from New York Botanical Garden (NYBG) of unknown origin. It is worth noting that *S. actinophylla* is widespread across the world as an ornamental primarily from Australian stock, which suggests this NYBG collection might be Australian in origin. Interestingly, *S. thaumasiantha* is also cultivated locally in SE New Guinea (Frodin, 1975), which could help explain the observed morphological similarities. Previous work on domestication points to possible multiple origins in a number of crops, with parallelism and convergence being the norm (Fuller et al., 2014; Purugganan, 2019).

Papuoschefflera s.l. was reconstructed as either paraphyletic with respect to Brassaia (**Figure 3** right) or as reciprocally monophyletic and sister to Brassaia with some (**Figure 4** left) or no support (**Figure 5** right). Although we could not place Cephaloschefflera (represented by *S. eriocephala*) with confidence, we found that morphogroup Pagiophyllae (represented by *S.* "*frigidariorum*" sp. ined.) belonged in Papuoschefflera s.s. (see section Phylogenetic Relationships in *Schefflera* in Results), together with morphogroups Bougainvilleanae, Schumannianae, and Versteegiae (**Figures 3–5**). Other Papuoschefflera s.l. morphogroups (e.g., Ischyrocephalae or Oreopolae) were monophyletic in some but not all of our trees, which should be further explored.

Evolutionary History of Papuasian Lineages

The sampled accessions of Papuasian *Schefflera* tended to form ecological and morphological clades within broader geographical clades. Analogous patterns have been observed in other Araliaceae clades, such as *Polyscias* (Plunkett and Lowry, 2010), Neotropical *Schefflera* (Fiaschi and Plunkett, 2011) and *Plerandra*, which is the Melanesian clade of *Schefflera* s.l. (Plunkett and Lowry, 2012). The dominance of Brassaia and Papuoschefflera s.l., on either side of the New Guinea

TABLE 4 | Morphological characters distinguishing Brassaia and Papuoschefflera s.l.

Morphological character	Brassaia	Papuoschefflera s.l.
Ovary position	1/2–2/3 superior	<1/6 superior
Floral bracts, ligules	Glabrous	Setulose
Main inflorescence axis	Sessile/short	Up to 60 cm long
Flower color	Red to pink	Pale green to white, some purple to red

Highlands also recalls a similar arrangement in the two clades of Afro-Malagasy *Schefflera* s.l. on either side of the Mozambique Channel, which are now recognized as genera *Astropanax* and *Neocussonia* (Gostel et al., 2017). These distribution patterns may prove to be fertile ground for the testing of biogeographic hypotheses.

The Woodlark Plate: A Source Area for Papuasian *Schefflera*

Though our estimated crown age of ∼26.3 Ma for the Papuasian *Schefflera* clade is older than the crown age of ∼21.9 Ma estimated by Li and Wen (2014) – which is the source of our secondary time constraints –, as expected, the highest posterior density intervals for both these estimates overlap. These divergence dates may seem to be at odds with the assertion by Lohman et al. (2011) that most of New Guinea was submerged prior to 20 Ma. Yet, several studies have also inferred the origin of Papuasian taxa back to the Oligocene (Cibois et al., 2014; Kodandaramaiah et al., 2018) or even as early as the late Eocene (Jønsson et al., 2011; Cozzarolo et al., 2019). During the Oligocene, there may have been an island archipelago, located where present-day New Guinea is, formed by the collision of both the Philippine-Sea and the Pacific plates with the Australia plate (Hill and Hall, 2003). The largest landmass would probably have been where the present-day Papuan Peninsula is and have resulted from the docking of the East Papua Composite Terrane (EPCT; part of the Woodlark plate) onto the Australia plate (Davies, 2012). Stratigraphic examination of sediments deposited in the Aure Trough provides evidence for mountain building on the Woodlark plate during this period (van Ufford and Cloos, 2005), indicating that the terranes forming the Papuan Peninsula were already emergent. Our reconstruction of the Woodlark plate as the best-supported ancestral area for Papuasian *Schefflera* (**Figure 5**) is consistent with the above scenario.

Hall (2012) estimated that the Sahul and Sunda shelves first collided ∼25 Ma (see Figure 32 in Hall, 2012). Phylogenies reconstructed by Li and Wen (2014) and Plunkett et al. (2005) and this study support Papuasian *Schefflera* as nested within Sundan Heptapleurum. It is thus within reason that our ancestral area reconstruction suggests that the common ancestor of Papuasian *Schefflera* dispersed from Sunda to colonize the Woodlark plate right before this contact took place (**Figure 5**). In this scenario, we hypothesize that a light-loving and pioneer ancestral Papuasian *Schefflera* would have rapidly colonized areas of the New Guinea land mass as they gradually emerged above sea level. The Woodlark plate could, therefore, have functioned as a source area for the colonization of New Guinea along the predominantly west-to-east axis of the Sunda-Sahul floristic exchange (Crayn et al., 2015).

Vicariance in Brassaia

Our reconstruction of the Sahul shelf as the ancestral area for Brassaia points to a vicariance scenario for the early evolutionary history of this clade (**Figure 5**). The Brassaia crown is dated at ∼18.1 Ma, which is earlier than the ∼5 Ma date estimated for the emergence of the Sahul shelf to form the southern half of New Guinea (Hall, 2009). As Brassaia also occurs on the southern fall of the Owen Stanley Range (near present-day Port Moresby, in the "tail" of Papua New Guinea) and is partly formed by the Sahul shelf, it is possible that ancestral Brassaia evolved in isolation from ancestral Papuoschefflera s.l. on either side of the proto-Owen Stanley Range during the early Miocene.

The placement of *S. kraemeri* in Brassaia supports the morphological circumscription of the group, despite this species' disjunct distribution with regard to the rest of the group. *Schefflera kraemeri* is found only in the Truk Islands, more than 800 km away from Papuasia. It is most likely to have arrived via long distance dispersal as there are no intervening landmasses in the Pacific Ocean to serve as stepping-stones. Human-mediated dispersal is unlikely, not only because of the inferred timing (which predates humanity), but also because *Schefflera* has limited uses in Papuasia and *S. kraemeri* has not been recorded among the region's indigenous people as a useful species (Cámara-Leret and Dennehy, 2019a,b). The estimated divergence time of ∼6.8 Ma is consistent with the geological age of these islands, which were determined to have been the result of volcanism ∼11 Ma (Keating et al., 1984).

Papuoschefflera s.l. Speciated on Geographic and "Sky" Islands

The divergence of the major clades of Papuoschefflera s.l. between ∼24.6 and 16.7 Ma overlaps with the period in the early Miocene when most of New Guinea was submerged (**Figure 5**). While Ischyrocephalae and the Eastern clade are inferred to have remained on the Woodlark plate at this time, Papuoschefflera s.s. may have arisen from an early dispersal to then-emerged islands in the Maoke plate, corresponding to the present-day Maoke Mountains. The splitting of the Eastern clade into Cephaloschefflera and the Oreopolae + Barbatae clade probably resulted from another dispersal to Sunda shelf terranes between ∼17.3 and 12.6 Ma. Indeed, zoochorous dispersal across narrow water barriers has been found to play an important role in the intercontinental floristic exchange of the Malesian flora (Crayn et al., 2015). The Papuasian *Schefflera*, with their fleshy drupaceous fruits, could have been widely dispersed (for example by birds; Snow, 1981) across the proto-Papuasian archipelago.

Accessions from islands located in the Bird's Head plate illustrate how *Schefflera* species may have colonized islands in geological times. The shallowest seafloor connecting the islands of Biak and Waigeo to mainland New Guinea is about 200 m deep, precluding an overland connection even during the Pleistocene, when sea level was up to 120 m shallower than nowadays (Voris, 2000). The divergence of *S.* "*bougainvill.biak*" sp. ined. ∼5.6 Ma coincides with the end of a 6-Myr upper Oligocene hiatus in deposition in the Biak Basin near Biak Island (Gold et al., 2014), indicating that sea level was shallower at that point in time. A larger area of land would have been exposed on both the island and the mainland, facilitating dispersal over a narrower water barrier. Similarly, the divergence of *Schefflera* sp. (Raja Ampat) ∼4.7 Ma coincided with an active period of Pliocene deformation, resulting in more land emerging above sea level (Charlton et al., 1991). The placement of this collection as sister to *S.* "*KB22*" (Sorong) sp. ined. suggests descent from a common ancestor that occupied the NW Bird's Head peninsula.

Ischyrocephalae is restricted to upper montane forests and subalpine grasslands, which could be explained in terms of phylogenetic biome conservatism – phenomenon that has been observed in vascular plants from the Malesian island of Borneo (Merckx et al., 2015) and also worldwide (Crisp et al., 2009). *Schefflera ischyrocephala* and *S. pilematophora* are found in the Saruwaged Range on the NE coast of New Guinea, which is interpreted to be a terrane of the South Bismarck plate. These taxa are inferred to have arrived separately from the Woodlark plate sometime between ∼11.6 and 10.7 Ma to the present, respectively, concurrent with the Finisterre volcanic arc accretion to the EPCT (Davies, 2012). This pattern could suggest that adaptation to montane regions in ancestral Papuoschefflera s.l. may have a role in promoting the highly diverse "sky island" flora of New Guinea (Sklenáø et al., 2014). Given the prevalence of high- and mid-elevation taxa both within the clade and its most recently diverging sister lineages, as well as a putative temperate origin for Asian *Schefflera* (Valcárcel and Wen, 2019), Papuasian *Schefflera* may have been pre-adapted to these environments (Antonelli, 2015). Pre-adaptation has also been invoked in a broad range of high-elevation taxa on Gunung Kinabalu (Merckx et al., 2015), a Malesian mountain located in Borneo, with uplift timing similar to that of the New Guinea Highlands. In both cases, however, this hypothesis remains to be tested by ancestral trait reconstruction, while accounting for diversification rate shifts.

Based on morphological characters, Philipson (1978) mused that *Schefflera* may be "vigorously diversifying at the present time." In Papuasia, the species richness of various plant and animal taxa has been shown to peak at mid-elevations (Colwell et al., 2016; Vollering et al., 2016). Papuasian *Schefflera* exhibits a similar distribution in species richness, which may have arisen from rapid adaptive radiation into new ecological niches created by mountain-building. This evolutionary mechanism has also been observed in the genus *Pseuduvaria* (Annonaceae) on New Guinea (Su and Saunders, 2009), various genera of Andean alpine plants (Nürk et al., 2018), and Ericaceae in montane regions worldwide (Schwery et al., 2015). Many biotic and environmental factors, including its presumed temperate ancestry and the ready availability of new uncolonized habitats resulting from the uplift of the New Guinea Highlands, would have predisposed Papuasian *Schefflera* to speciation via biome shifts (Donoghue and Edwards, 2014). Our chronogram hints to a speciation burst in the last 12–5 Myr and we hypothesize recent adaptive radiation, facilitated by mountain building, could be the primary driver for the present-day diversity of *Schefflera*. Further insights into this phenomenon could be achieved by examining an expanded sample of taxa for trait shifts and diversification rates along an elevational gradient.

CONCLUSION

Several phylogeographical studies on Papuasian fauna have produced comparable results that point to major themes in Papuasian biogeography at different points in geological time that we recapitulate below.

Woodlark plate: source area for the colonization of New Guinea in the late Oligocene. Our inference that the Woodlark plate acted as the source area for the colonization of Papuasia by *Schefflera* in the late Oligocene is supported by faunistic studies on the endemic frog genus *Mantophryne* (Oliver et al., 2013). A similar pattern of colonization, albeit from Australia, would also explain the present-day distribution of Gondwana-derived Melanotaeniid rainbowfish (Unmack et al., 2013). However, more studies involving divergence time estimation and ancestral area reconstruction are required to establish the role of the Woodlark plate in the diversification of Papuasian taxa.

New Guinea Highlands: a barrier since the late Oligocene. The divergence of Papuasian *Schefflera* into northern Papuoschefflera and southern Brassaia – along a topographical barrier running east to west across the length of Papuasia – lends support to the role played by the New Guinea Highlands with respect to north-south disjunctions and has also been observed in lineages of freshwater organisms such as Melanotaeniid rainbowfish (Unmack et al., 2013), *Mantophryne* frogs (Oliver et al., 2013), and the turtle *Elseya novoguinae* (Georges et al., 2014). The more recent rapid uplift of the New Guinea Highlands in the late Miocene and Pliocene epochs coincides with the divergence of *Sericulus* bowerbird species (Zwiers et al., 2008) and populations of the passerine bird-species *Colluricincla megarhyncha* (Deiner et al., 2011). Thus, it is possible that initial mountain-building created physical barriers by compartmentalizing the Papuasian terranes into separate basins. Subsequently, the Pliocene uplift would then have raised the New Guinean Highlands to the point that new ecological barriers (i.e., alpine and subalpine zones) effectively isolated populations adapted to lower elevations.

Geographic and ecological ("sky") islands shaped evolutionary relationships, both deep and shallow. In the Miocene, when Papuasia existed only as a chain of islands, major clades nested in Papuoschefflera s.l. originated and diversified into high- and mid-elevation clades across New Guinea's mountain ranges. Speciation on geographic islands is evident in the Miocene divergence of freshwater taxa on the Bird's Head Peninsula and Maoke terranes (Unmack et al., 2013; Georges et al., 2014), which probably existed as isolated landmasses prior to docking with the Maoke plate. These putative geographic islands have also been cited as cradles of diversity for corvoid birds (Jønsson et al., 2011) and *Ptilinopus* fruit doves (Cibois et al., 2014). Similarly, *Orthonyx* logrunners in the Bird's Head Peninsula have been shown to be genetically distinct from those in the rest of New Guinea (Joseph et al., 2001). Additionally, a more recent genetic divergence (late Miocene onward) has been found on an unnamed *Petaurus* glider species on Normanby Island (Malekian et al., 2010). Isolation on an ecological island in the Cromwell Range during the Pleistocene may also explain the observed genetic differentiation in an isolated population of the pademelon *Thylogale browni browni*, which is restricted to the edges of subalpine forests (Macqueen et al., 2011).

New ecological niches followed the New Guinea Highlands uplift and could have driven rapid recent radiations. The divergence dates of most Papuasian *Schefflera* taxa in our chronogram coincide with the uplift of the New Guinea Highlands in the late Miocene. This geological event created

new ecological niches and has been invoked as the primary driver for the diversity of several groups of Papuasian mammals and birds, such as Australasian rats (Rowe et al., 2011), *Dendrolagus* tree kangaroos (Eldridge et al., 2018), *Exocelina* diving beetles (Toussaint et al., 2014), *Meliphaga* honeyeaters (Norman et al., 2007), *Prenolepis* ants (Matos-Maraví et al., 2018), Pseudocheiridae ringtail possums (Meredith et al., 2010), *Syma* kingfishers (Linck et al., 2019), *Thraulus* mayflies (Cozzarolo et al., 2019), and *Thylogale* pademelons (Macqueen et al., 2011). Together, these studies strongly suggest that rapid adaptive radiation into newly created ecological niches resulting from recent mountain-building would best explain the richness of Papuasia's biodiversity.

In closing, we have found Asian *Schefflera* to be among the first plant genera to have crossed from Sunda toward the Australian plate at the start of the Sunda-Sahul floristic exchange in the late Oligocene. The widespread distribution of this lineage and its existence in Papuasia since its known geologic origin suggest that its evolutionary history will prove instructive in understanding the region's plant diversity. Our study demonstrates that Hyb-Seq with universal probes on a sample set comprising mostly herbarium specimens can resolve both deep and shallow phylogenetic relationships to elucidate the drivers of this diversity. Our results suggest an important role for (1) the Woodlark plate (present-day Papuan Peninsula), (2) the New Guinea Highlands, (3) isolation on geographic and "sky" islands, and (4) the late Miocene New Guinea Highlands uplift in explaining plant biogeography in Papuasia.

DATA AVAILABILITY STATEMENT

The raw reads (FASTQ files) are available from the NCBI BioProject database (ID: PRJNA604390). The datasets generated and analyzed are available from Zenodo (doi: 10.5281/zenodo.3534088).

AUTHOR CONTRIBUTIONS

The study and sampling scheme were jointly conceived by all authors based on taxonomic expertise and collection data provided by DF. ZS and LP sampled the specimens and carried out the molecular work and data analysis. ZS wrote the manuscript, with contributions from all authors.

FUNDING

LP had funding from the Garfield Weston Foundation (Global Tree Seed Bank Project) and EU-SYNTHESYS (NL-TAF-6894). RC-L had funding from EU-SYNTHESYS (GB-TAF-6305). This article was based on a final manuscript by ZS for his MSc in Plant and Fungal Taxonomy, Diversity and Conservation at Queen Mary, University of London and at RBG Kew, funded by a scholarship from the National Parks Board, Singapore.

ACKNOWLEDGMENTS

We thank our colleagues at Royal Botanic Gardens, Kew for their invaluable assistance: Edith Kapinos, Isabel Fairlie, Juan Viruel, Laszlo Csiba, Olivier Maurin, Oscar Pérez-Escobar, Marcelo Sellaro, Miranda Janatka, Penny Malakasi, Robyn Cowan, Sally Dawson, Sidonie Bellot, Tim Utteridge, Bill Baker, Félix Forest, and Ilia Leitch. We extend our gratitude to the herbaria of K, L, BM, SING and to Julian Schrader (GAU Göttingen) for providing samples, and to GenBank, 1KP, and PAFTOL for making sequences publicly available. We thank the Sackler Phylogenomic Laboratory and the Jodrell Laboratory at RBG Kew for facilitating molecular lab resources and computing resources.

DEDICATION

We dedicate this manuscript to the memory of our friend and mentor David Gamman Frodin, who passed away during the course of this study (1940–2019). His legacy lives on in the "*Schefflera* Team" and the scientists he inspired during his fruitful career.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00258/full#supplementary-material

REFERENCES

Andrews, S. (2018). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed August 9, 2019).

Antonelli, A. (2015). Biodiversity: multiple origins of mountain life. *Nature* 524, 300–301. doi: 10.1038/nature14645

Bacon, C. D., Michonneau, F., Henderson, A. J., McKenna, M. J., Milroy, A. M., and Simmons, M. P. (2013). Geographic and taxonomic disparities in species diversity: dispersal and diversification rates across Wallace's line. *Evolution* 67, 2058–2071. doi: 10.1111/evo.12084

Baele, G., Lemey, P., and Vansteelandt, S. (2013). Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. *BMC Bioinformatics* 14:85. doi: 10.1186/1471-2105-14-85

Baldwin, S. L., Fitzgerald, P. G., and Webb, L. E. (2012). Tectonics of the New Guinea Region. *Annu. Rev. Earth Planet. Sci.* 40, 495–520. doi: 10.1146/annurev-earth-040809-152540

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. Computat. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Beehler, B. M. (2007). "Papuan terrestrial biogeography, with special reference to birds," in *The Ecology of Papua: Part One*, eds A. J.

Marshall, and B. M. Beehler (Singapore: Periplus Editions (HK) Ltd), 196–206.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., et al. (2018). GenBank. *Nucleic Acids Res.* 46, D41–D47. doi: 10.1093/nar/gkx1094

Bentham, G. (1867). "Araliaceae," in *Genera Plantarum*, eds G. Bentham, and J. D. Hooker (London: William Pamplin), 931–941.

Bird, P. (2003). An updated digital model of plate boundaries. *Geochem. Geophys. Geosyst.* 4, 13–50. doi: 10.1029/2001GC000252

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* 4:e1660. doi: 10.7717/peerj.1660

Brass, L. J. (1941). The 1938-39 expedition to the snow mountains, Netherlands New Guinea. *J. Arnold Arbor.* 22, 271–295.

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

Brummitt, R. K. (ed.) (2001). "World geographical scheme for recording plant distributions," in *Hunt Institute for Botanical Documentation*, 2nd Edn (Pittsburgh: Carnegie Mellon University).

Cámara-Leret, R., and Dennehy, Z. (2019a). Indigenous knowledge of New Guinea's useful plants: a review. *Econ. Bot.* 73, 405–415. doi: 10.1007/s12231-019-09464-1

Cámara-Leret, R., and Dennehy, Z. (2019b). Information gaps in indigenous and local knowledge for science-policy assessments. *Nat. Sustain.* 2, 736–741. doi: 10.1038/s41893-019-0324-0

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

Charlton, T. R., Hall, R., and Partoyo, R. (1991). The geology and tectonic evolution of Waigeo Island, NE Indonesia. *J. Southeast Asian Earth Sci.* 6, 289–297. doi: 10.1016/0743-9547(91)90074-8

Chau, J. H., Rahfeldt, W. A., and Olmstead, R. G. (2018). Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Appl. Plant Sci.* 6:e1032. doi: 10.1002/aps3.1032

Cibois, A., Thibault, J.-C., Bonillo, C., Filardi, C. E., Watling, D., and Pasquet, E. (2014). Phylogeny and biogeography of the fruit doves (Aves: Columbidae). *Mol. Phylogenet. Evol.* 70, 442–453. doi: 10.1016/j.ympev.2013.08.019

Colwell, R. K., Gotelli, N. J., Ashton, L. A., Beck, J., Brehm, G., Fayle, T. M., et al. (2016). Midpoint attractors and species richness: modelling the interaction between environmental drivers and geometric constraints. *Ecol. Lett.* 19, 1009–1022. doi: 10.1111/ele.12640

Cozzarolo, C.-S., Balke, M., Buerki, S., Arrigo, N., Pitteloud, C., Gueuning, M., et al. (2019). Biogeography and ecological diversification of a Mayfly Clade in New Guinea. *Front. Ecol. Evol.* 7:233. doi: 10.3389/fevo.2019.00233

Crayn, D. M., Costion, C., and Harrington, M. G. (2015). The Sahul-Sunda floristic exchange: dated molecular phylogenies document cenozoic intercontinental dispersal dynamics. *J. Biogeogr.* 42, 11–24. doi: 10.1111/jbi.12405

Crisp, M. D., Arroyo, M. T. K., Cook, L. G., Gandolfo, M. A., Jordan, G. J., McGlone, M. S., et al. (2009). Phylogenetic Biome Conservatism on a Global Scale. *Nature* 458, 754–756. doi: 10.1038/nature07764

Crisp, M. D., Trewick, S. A., and Cook, L. G. (2011). Hypothesis testing in biogeography. *Trends Ecol. Evol.* 26, 66–72. doi: 10.1016/j.tree.2010.11.005

Davies, H. L. (2012). The geology of New Guinea - the Cordilleran Margin of the Australian Continent. *Episodes* 35, 1–16.

Deiner, K., Lemmon, A. R., Mack, A. L., Fleischer, R. C., and Dumbacher, J. P. (2011). A passerine Bird's evolution corroborates the geologic history of the Island of New Guinea. *PLoS One* 6:e19479. doi: 10.1371/journal.pone.0019479

Der Sarkissian, C., Allentoft, M. E., Ávila-Arcos, M. C., Barnett, R., Campos, P. F., Cappellini, E., et al. (2015). Ancient Genomics. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20130387. doi: 10.1098/rstb.2013.0387

Dodsworth, S., Pokorny, L., Johnson, M. G., Kim, J. T., Maurin, O., Wickett, N. J., et al. (2019). Hyb-Seq for flowering plant systematics. *Trends Plant Sci.* 24, 887–891. doi: 10.1016/j.tplants.2019.07.011

Donoghue, M. J., and Edwards, E. J. (2014). Biome shifts and niche evolution in plants. *Annu. Rev. Ecol. Evol. Syst.* 45, 547–572. doi: 10.1146/annurev-ecolsys-120213-091905

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, 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. doi: 10.1093/sysbio/syr126

Eldridge, M. D. B., Potter, S., Helgen, K. M., Sinaga, M. H., Aplin, K. P., Flannery, T. F., et al. (2018). Phylogenetic analysis of the tree-kangaroos (*Dendrolagus*) reveals multiple divergent lineages within New Guinea. *Mol. Phylogenet. Evol.* 127, 589–599. doi: 10.1016/j.ympev.2018.05.030

Fiaschi, P., and Plunkett, G. M. (2011). Monophyly and phylogenetic relationships of neotropical *Schefflera* (Araliaceae) based on plastid and nuclear markers. *Syst. Bot.* 36, 806–817. doi: 10.1600/036364411X583754

Finch, K. N., Jones, F. A., and Cronn, R. C. (2019). Genomic resources for the neotropical tree genus *Cedrela* (Meliaceae) and its relatives. *BMC Genomics* 20:58. doi: 10.1186/s12864-018-5382-6

Forrest, L. L., Hart, M. L., Hughes, M., Wilson, H. P., Chung, K.-F., Tseng, Y. H., et al. (2019). The limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. *Front. Ecol. Evol.* 7:439. doi: 10.3389/fevo.2019.00439

Frodin, D. G. (1975). Studies in *Schefflera* (Araliaceae): the Cephaloschefflera complex. *J. Arnold Arbor.* 56, 427–448.

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

Frodin, D. G., and Govaerts, R. (2003). *World Checklist and Bibliography of Araliaceae*. Kew: Royal Botanic Gardens.

Frodin, D. G., Lowry, P. P. II, and Plunkett, G. M. (2010). *Schefflera* (Araliaceae): taxonomic history, overview and progress. *Plant Divers. Evol.* 128, 561–595. doi: 10.1127/1869-6155/2010/0128-0028

Fuller, D. Q., Denham, T., Arroyo-Kalin, M., Lucas, L., Stevens, C. J., Qin, L., et al. (2014). Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6147–6152. doi: 10.1073/pnas.1308937110

Georges, A., Zhang, X., Unmack, P., Reid, B. N., Le, M., and McCord, W. P. (2014). Contemporary genetic structure of an endemic freshwater turtle reflects miocene Orogenesis of New Guinea. *Biol. J. Linn. Soc.* 11, 192–208. doi: 10.1111/bij.12176

Gold, D., Hall, R., Burgess, P., and BouDagher-Fadel, M. (2014). "The Biak Basin and its setting in the bird's head region of West Papua," in *Proceedings of the Thirty-Eighth Annual Convention & Exhibition on Indonesian Petroleum Association, IPA14–G–298*, Jakarta.

Gostel, M. R., Plunkett, G. M., and Lowry, P. P. II (2017). Straddling the Mozambique Channel: molecular evidence for two major clades of Afro-Malagasy *Schefflera* (Araliaceae) co-occurring in Africa and Madagascar. *Plant Ecol. Evol.* 150, 87–108. doi: 10.5091/plecevo.2017.1193

Hall, R. (2009). Southeast Asia's changing Palaeogeography. *Blumea* 54, 148–161. doi: 10.3767/000651909X475941

Hall, R. (2012). Late Jurassic–Cenozoic reconstructions of the Indonesian Region and the Indian Ocean. *Tectonophysics* 57, 1–41. doi: 10.1016/j.tecto.2012.04.021

Harms, H. (1921). Die Araliaceae Papuasiens. *Bot. Jahrb. Syst. Pflanzengesch. Pflanzengeogr.* 56, 374–414.

Hart, M. L., Forrest, L. L., Nicholls, J. A., and Kidner, C. A. (2016). Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 65, 1081–1092. doi: 10.12705/655.9

Hartmann, S., and Vision, T. J. (2008). Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol. Biol.* 8:95. doi: 10.1186/1471-2148-8-95

Heads, M. (2003). Ericaceae in Malesia: vicariance biogeography, terrane tectonics and ecology. *Telopea* 10, 311–449. doi: 10.3732/ajb.0900109

Hill, K. C., and Hall, R. (2003). "Mesozoic-cenozoic evolution of Australia's New Guinea margin in a West Pacific context," in *Evolution and Dynamics of the Australian Plate*, eds R. R. Hillis, and R. D. Müller (Sydney: Geological Society of Australia), 265–290.

Ho, S. Y. W., Phillips, M. J., Cooper, A., and Drummond, A. J. (2005). Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.* 22, 1561–1568. doi: 10.1093/molbev/msi145

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

Hoover, J. D., Kumar, S., James, S. A., Leisz, S. J., and Laituri, M. (2017). Modeling hotspots of plant diversity in New Guinea. *Trop. Ecol.* 58, 623–640.

Johns, R. J., Shea, G. A., and Puradyatmika, P. (2007a). "Subalpine and Alpine Vegetation of Papua," in *The Ecology of Papua: Part Two*, eds A. J. Marshall, and B. M. Beehler (Singapore: Periplus Editions (HK) Ltd), 1025–1053.

Johns, R. J., Shea, G. A., Vink, W., and Puradyatmika, P. (2007b). "Montane vegetation of Papua," in *The Ecology of Papua: Part Two*, eds A. J. Marshall, and B. M. Beehler (Singapore: Periplus Editions (HK) Ltd), 977–1024.

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. (2018). 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

Jønsson, K. A., Fabre, P.-H., Ricklefs, R. E., and Fjeldsa, J. (2011). Major global radiation of corvoid birds originated in the proto-papuan archipelago. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2328–2333. doi: 10.1073/pnas.1018956108

Joseph, L., Slikas, B., Alpers, D., and Schodde, R. (2001). Molecular systematics and phylogeography of New Guinean Logrunners (Orthonychidae). *Emu* 101, 273–280. doi: 10.1071/MU01008

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

Kadlec, M., Bellstedt, D. U., Le Maitre, N. C., and Pirie, M. D. (2017). Targeted NGS for species level phylogenomics: "made to measure" or "one size fits all"? *PeerJ* 5:e3569. doi: 10.7717/peerj.3569

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

Kates, H. R., Johnson, M. G., Gardner, E. M., Zerega, N. J. C., and Wickett, N. J. (2018). Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. *Am. J. Bot.* 105, 404–416. doi: 10.1002/ajb2.1068

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

Keating, B. H., Mattey, D. P., Helsley, C. E., Naughton, J. J., Epp, D., Lazarewicz, A., et al. (1984). Evidence for a hot spot origin of the Caroline Islands. *J. Geophysical Res. Solid Earth* 89, 9937–9948. doi: 10.1029/JB089iB12p09937

Kim, K., and Yang, T. -J. (2016). *Data from: The Complete Chloroplast Genome of Aralia elata. GenBank*. Available at: https://www.ncbi.nlm.nih.gov/nuccore/KT153023.1 (accessed August 9, 2019).

Kodandaramaiah, U., Braby, M. F., Grund, R., Müller, C. J., and Wahlberg, N. (2018). Phylogenetic relationships, biogeography and diversification of Coenonymphina butterflies (Nymphalidae: Satyrinae): intercontinental dispersal of a Southern Gondwanan Group? *Syst. Entomol.* 43, 798–809. doi: 10.1111/syen.12303

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, B. N. (2014). *Solid Phase Reverse Immobilization (SPRI) Bead Technology for Micro RNA Clean up Using the Agencourt RNAclean XP Kit*. Ph.D. thesis, Beckman-Coulter, Brea, CA.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5:e1000520. doi: 10.1371/journal.pcbi.1000520

Lemmon, E. M., and Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Ann. Rev. Ecol. Evol. Syst.* 44, 99–121. doi: 10.1146/annurev-ecolsys-110512-135822

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/1303.3997

Li, R., Ma, P.-F., Wen, J., and Yi, T.-S. (2013). Complete sequencing of five Araliaceae chloroplast genomes and the phylogenetic implications. *PLoS One* 8:e78568. doi: 10.1371/journal.pone.0078568

Li, R., and Wen, J. (2014). Phylogeny and biogeography of Asian *Schefflera* (Araliaceae) based on nuclear and plastid DNA sequence data. *J. Syst. Evol.* 52, 431–449. doi: 10.1111/jse.12052

Linck, E., Freeman, B. G., and Dumbacher, J. P. (2019). Speciation with gene flow across an elevational gradient in New Guinea Kingfishers. *bioRxiv* [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/589044v1

Liu, Y., Johnson, M. G., Cox, C. J., Medina, R., Devos, N., Vanderpoorten, A., et al. (2019). Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. *Nat. Commun.* 10:1485. doi: 10.1038/s41467-019-09454-w

Lohman, D. J., de Bruyn, M., Page, T., von Rintelen, K., Hall, R., Ng, P. K. L., et al. (2011). Biogeography of the Indo-Australian Archipelago. *Annu. Rev. Ecol. Evol. Syst.* 42, 205–226. doi: 10.1146/annurev-ecolsys-102710-145001

Lowry, P. P. II, and Plunkett, G. M. (2010). Recircumscription of Polyscias (Araliaceae) to include six related genera, with a new infrageneric classification and a synopsis of species. *Plant Divers. Evol.* 128, 55–84. doi: 10.1127/1869-6155/2010/0128-0003

Macqueen, P., Goldizen, A. W., Austin, J. J., and Seddon, J. M. (2011). Phylogeography of the Pademelons (Marsupialia: Macropodidae: *Thylogale*) in New Guinea Reflects Both Geological and Climatic Events during the Plio-Pleistocene. *J. Biogeogr.* 38, 1732–1747. doi: 10.1111/j.1365-2699.2011.02522.x

Mai, U., and Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* 19:272. doi: 10.1186/s12864-018-4620-2

Mairal, M., Pokorny, L., Aldasoro, J. J., Alarcón, M., and Sanmartín, S. (2015). Ancient vicariance and climate-driven extinction explain continental-wide disjunctions in Africa: the case of the Rand Flora genus *Canarina* (Campanulaceae). *Mol. Ecol.* 24, 1335–1354. doi: 10.1111/mec.13114

Mairal, M., Sanmartín, I., Herrero, A., Pokorny, L., Vargas, P., Aldasoro, J. J., et al. (2017). Geographic barriers and Pleistocene climate change shaped patterns of genetic variation in the Eastern Afromontane biodiversity hotspot. *Sci. Rep.* 7:45749. doi: 10.1038/srep45749

Malekian, M., Cooper, S. J. B., Norman, J. A., Christidis, L., and Carthew, S. M. (2010). Molecular systematics and evolutionary origins of the genus *Petaurus* (Marsupialia: Petauridae) in Australia and New Guinea. *Mol. Phylogenet. Evol.* 54, 122–135. doi: 10.1016/j.ympev.2009.07.026

Marshall, A. J. (2007). "The diversity and conservation of Papua's ecosystems," in *The Ecology of Papua: Part Two*, eds A. J. Marshall, and B. M. Beehler (Singapore: Periplus Editions (HK) Ltd), 753–770.

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., et al. (2014). Data access for the 1,000 Plants (1KP) Project. *Gigascience* 3:17. doi: 10.1186/2047-217X-3-17

Matos-Maraví, P., Clouse, R. M., Sarnat, E. M., Economo, E. P., LaPolla, J. S., Borovanska, M., et al. (2018). An Ant Genus-Group (*Prenolepis*) Illuminates the Biogeography and Drivers of Insect Diversification in the Indo-Pacific. *Mol. Phylogenet. Evol.* 123, 16–25. doi: 10.1016/j.ympev.2018.02.007

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

Meiklejohn, K. A., Faircloth, B. C., Glenn, T. C., Kimball, R. T., and Braun, E. L. (2016). Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. *Syst. Biol.* 65, 612–627. doi: 10.1093/sysbio/syw014

Merckx, V. S. F. T., Hendriks, K. P., Beentjes, K. K., Mennes, C. B., Becking, L. E., Peijnenburg, K. T. C. A., et al. (2015). Evolution of endemism on a young tropical mountain. *Nature* 524, 347–350. doi: 10.1038/nature14949

Meredith, R. W., Mendoza, M. A., Roberts, K. K., Westerman, M., and Springer, M. S. (2010). A Phylogeny and timescale for the evolution of Pseudocheiridae (Marsupialia: Diprotodontia) in Australia and New Guinea. *J. Mamm. Evol.* 17, 75–99. doi: 10.1007/s10914-010-9129-7

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. (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

Mirarab, S. (2019). Species tree estimation using ASTRAL: practical considerations. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/1904.03826v2

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., and Warnow, T. (2015). PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J. Comput. Biol.* 22, 377–386. doi: 10.1089/cmb.2014.0156

Murphy, B., Forest, F., Barraclough, T., Rosindell, J., Bellot, S., Cowan, R., et al. (2020). A phylogenomic analysis of *Nepenthes* (Nepenthaceae). *Mol. Phylogenet. Evol.* 144:1066682. doi: 10.1016/j.ympev.2019.106668

Mutke, J., Sommer, J. H., Kreft, H., Kier, G., and Barthlott, W. (2011). "Vascular plant diversity in a changing world: global centres and biome-specific patterns," in *Biodiversity Hotspots: Distribution and Protection of Conservation Priority Areas*, eds F. E. Zachos, and J. C. Habel (Heidelberg: Springer), 83–96.

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

Nguyen, N.-P. D., Mirarab, S., Kumar, K., and Warnow, T. (2015). Ultra-large alignments using phylogeny-aware profiles. *Genome Biol.* 16:124. doi: 10.1186/s13059-015-0688-z

Nicolas, A. N., and Plunkett, G. M. (2014). Diversification times and biogeographic patterns in Apiales. *Bot. Rev.* 80, 30–58. doi: 10.1007/s12229-014-9132-4

Norman, J. A., Rheindt, F. E., Rowe, D. L., and Christidis, L. (2007). Speciation dynamics in the Australo-Papuan *Meliphaga* honeyeaters. *Mol. Phylogenet. Evol.* 42, 80–91. doi: 10.1016/j.ympev.2006.05.032

Nürk, N. M., Michling, F., and Linder, H. P. (2018). Are the radiations of temperate lineages in tropical alpine ecosystems pre-adapted? *Glob. Ecol. Biogeogr.* 27, 334–345. doi: 10.1111/geb.12699

Oliver, L. A., Rittmeyer, E. N., Kraus, F., Richards, S. J., and Austin, C. C. (2013). Phylogeny and phylogeography of *Mantophryne* (Anura: Microhylidae) reveals cryptic diversity in New Guinea. *Mol. Phylogenet. Evol.* 67, 600–607. doi: 10.1016/j.ympev.2013.02.023

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

Paijmans, K. (1976). "Vegetation," in *New Guinea Vegetation*, ed. K. Paijmans (Canberra: Australian National University Press), 23–105.

Philipson, W. R. (1978). "Araliaceae: growth forms and shoot morphology," in *Tropical Trees as Living Systems*, eds P. B. Tomlinson, and M. H. Zimmermann (Cambridge: Cambridge University Press), 269–284.

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *J. Theor. Biol.* 13, 131–144. doi: 10.1016/0022-5193(66)90013-0

Plunkett, G. M., and Lowry, P. P. II (2010). Paraphyly and Polyphyly in *Polyscias* Sensu Lato: molecular evidence and the case for Recircumscribing the "Pinnate Genera" of Araliaceae. *Plant Divers. Evol.* 128, 23–54. doi: 10.1127/1869-6155/2010/0128-0002

Plunkett, G. M., and Lowry, P. P. II (2012). Phylogeny and diversification in the melanesian *Schefflera* Clade (Araliaceae) based on evidence from nuclear rDNA spacers'. *Syst. Bot.* 37, 279–291. doi: 10.1600/036364412X616837

Plunkett, G. M., Lowry, P. P. II, Frodin, D. G., and Wen, J. (2005). Phylogeny and geography of *Schefflera*: pervasive polyphyly in the largest genus of Araliaceae. *Ann. Mo. Bot. Gard.* 92, 202–204.

Plunkett, G. M., Wen, J., and Lowry, P. P. II (2004). Infrafamilial classifications and characters in Araliaceae: insights from the phylogenetic analysis of nuclear (ITS) and Plastid (*trn*L-*trn*F) sequence data. *Plant Syst. Evol.* 245, 1–39. doi: 10.1007/s00606-003-0101-3

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

Purugganan, M. D. (2019). evolutionary insights into the nature of plant domestication. *Curr. Biol.* 29, R705–R714. doi: 10.1016/j.cub.2019.05.053

R Core Team, (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rambaut, A. (2018). *FigTree v1.4.4*. Available at: https://github.com/rambaut/figtree/releases/tag/v1.4.4 (accessed August 9, 2019).

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

Ravi, V., Khurana, J. P., Tyagi, A. K., and Khurana, P. (2008). An update on chloroplast genomes. *Plant Syst. Evol.* 271, 101–122. doi: 10.1007/s00606-007-0608-0

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

Rowe, K. C., Aplin, K. P., Baverstock, P. R., and Moritz, C. (2011). Recent and rapid speciation with limited morphological disparity in the genus *Rattus*. *Syst. Biol.* 60, 188–203. doi: 10.1093/sysbio/syq092

Sanmartín, I., van der Mark, P., and Ronquist, F. (2008). Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary Islands. *J. Biogeogr.* 35, 428–449. doi: 10.1111/j.1365-2699.2008.01885.x

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

Sayyari, E., and Mirarab, S. (2018). Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes* 9:E132. doi: 10.3390/genes9030132

Schwery, O., Onstein, R. E., Bouchenak-Khelladi, Y., Xing, Y., Carter, R. J., and Linder, H. P. (2015). As old as the mountains: the radiations of the Ericaceae. *New Phytol.* 207, 355–367. doi: 10.1111/nph.13234

Shen, X.-X., Salichos, L., and Rokas, A. (2016). A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. *Genome Biol. Evol.* 2565–2580. doi: 10.1093/gbe/evw179

Sklenáø, P., Hedberg, I., and Cleef, A. M. (2014). Island biogeography of tropical Alpine floras. *J. Biogeogr.* 41, 287–297. doi: 10.1111/jbi.12212

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

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

Snow, D. W. (1981). Tropical frugivorous birds and their food plants: a world survey. *Biotropica* 13, 1–14. doi: 10.2307/2387865

Soto Gomez, M., Pokorny, L., Kantar, M. B., Forest, F., Leitch, I. J., Gravendeel, B., et al. (2019). A customized nuclear target enrichment approach for developing a phylogenomic Baseline for Yams (Dioscoreaceae). *Appl. Plant Sci.* 7:e11254. doi: 10.1002/aps3.11254

Springer, M. S., Murphy, W. J., and Roca, A. L. (2018). Appropriate fossil calibrations and tree constraints uphold the Mesozoic divergence of Solenodons from other extant mammals. *Mol. Phylogenet. Evol.* 121, 158–165. doi: 10.1016/j.ympev.2018.01.007

Stadler, T. (2009). On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J. Theor. Biol.* 261, 58–66. doi: 10.1016/j.jtbi.2009.07.018

Su, Y. C. F., and Saunders, R. M. K. (2009). Evolutionary divergence times in the Annonaceae: evidence of a late Miocene origin of *Pseuduvaria* in Sundaland with subsequent diversification in New Guinea. *BMC Evol. Biol.* 9:153. doi: 10.1186/1471-2148-9-153

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 Evol.* 4:vey016. doi: 10.1093/ve/vey016

Toussaint, E. F., Hall, R., Monaghan, M. T., Sagata, K., Ibalim, S., Shaverdo, H. V., et al. (2014). The Towering Orogeny of New Guinea as a Trigger for Arthropod Megadiversity. *Nat. Commun.* 5:4001. doi: 10.1038/ncomms5001

Unmack, P. J., Allen, G. R., and Johnson, J. B. (2013). Phylogeny and Biogeography of Rainbowfishes (Melanotaeniidae) from Australia and New Guinea. *Mol. Phylogenet. Evol.* 67, 15–27. doi: 10.1016/j.ympev.2012.12.019

Valcárcel, V., and Wen, J. (2019). Chloroplast phylogenomic data support Paleocene - Eocene Amphi-pacific early radiation for the Asian Palmate Core Araliaceae. *J. Syst. Evol.* 57, 547–560. doi: 10.1111/jse.12522

Van Andel, T., Veltman, M. A., Bertin, A., Maat, H., Polime, T., Lambers, D., et al. (2019). Hidden rice diversity in the Guianas. *Front. Plant Sci.* 10:1161. doi: 10.3389/fpls.2019.01161

van Royen, P. (1979). *The Alpine Flora of New Guinea*. Vaduz: A. R. Gantner Verlag K. G.

van Ufford, A. Q., and Cloos, M. (2005). Cenozoic Tectonics of New Guinea. *AAPG Bull.* 89, 119–140. doi: 10.1306/08300403073

van Welzen, P. C., Parnell, J. A. N., and Ferry Slik, J. W. (2011). Wallace's line and plant distributions: Two or three phytogeographical areas and where to group java? *Biol. J. Linn. Soc.* 103, 531–545. doi: 10.1111/j.1095-8312.2011.01647.x

van Welzen, P. C., Turner, H., and Roos, M. (2001). New Guinea: a correlation between accreting areas and dispersing Sapindaceae. *Cladistics* 17, 242–247. doi: 10.1006/clad.2001.0173

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

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

Vollering, J., Schuiteman, A., de Vogel, E., van Vugt, R., and Raes, N. (2016). Phytogeography of New Guinean Orchids: patterns of species richness and turnover. *J. Biogeography* 43, 204–214. doi: 10.1111/jbi.12612

Voris, H. K. (2000). Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *J. Biogeogr.* 27, 1153–1167. doi: 10.1046/j.1365-2699.2000.00489.x

Wallace, A. R. (1869). *The Malay Archipelago: The Land of the Orang-Utan and the Bird of Paradise: A Narrative of Travel, with Studies of Man and Nature*. New York, NY: Harper & Brothers.

Warburg, O. (1891). Beiträge zur Kenntnis der papuanischen Flora. *Bot. Jahrb. Syst. Pflanzengesch. Pflanzengeogr.* 13, 230–455.

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, J., Ree, R. H., Ickert-Bond, S. M., Nie, Z., and Funk, V. (2013). Biogeography: Where do we go from here? *Taxon* 62, 912–927. doi: 10.12705/625.15

Wheeler, T. J., and Kececioglu, J. D. (2007). Multiple alignment by aligning alignments. *Bioinformatics* 23, i559–i568. doi: 10.1093/bioinformatics/btm226

White, D. M., Islam, M. B., and Mason-Gamer, R. J. (2019). Phylogenetic inference in section *Archerythroxylum* informs taxonomy, biogeography, and the domestication of coca (*Erythroxylum* Species). *Am. J. Bot.* 106, 154–165. doi: 10.1002/ajb2.1224

Wikramanayake, E. D. (2002). *Terrestrial Ecoregions of the Indo-Pacific: A Conservation Assessment*. Washington, DC: Island Press.

Williams, J. N. (2011). "Human population and the hotspots revisited: a 2010 assessment," in *Biodiversity Hotspots: Distribution and Protection of Conservation Priority Areas*, eds F. E. Zachos, and J. C. Habel (Heidelberg: Springer), 61–81.

Yule, G. U. (1925). A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F.R.S. *Philos. Trans. Roy. Soc. Lond. B* 213, 21–87.

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

Zong, X., Song, J., Lv, J., and Wang, S. (2016). The complete chloroplast genome sequence of *Schefflera octophylla*. *Mitochondrial DNA* 27, 4685–4686. doi: 10.3109/19401736.2015.1106502

Zwiers, P. B., Borgia, G., and Fleischer, R. C. (2008). Plumage based classification of the bowerbird genus *Sericulus* evaluated using a multi-gene, multi-genome analysis. *Mol. Phylogenet. Evol.* 46, 923–931. doi: 10.1016/j.ympev.2007.11.019

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: 17 June 2020
doi: 10.3389/fpls.2020.00798

Miocene Diversification in the Savannahs Precedes Tetraploid Rainforest Radiation in the African Tree Genus *Afzelia* (Detarioideae, Fabaceae)

Armel S. L. Donkpegan[1,2,3], Jean-Louis Doucet[1], Olivier J. Hardy[2], Myriam Heuertz[4†] and Rosalía Piñeiro[5,6†]*

[1] Forest is Life, TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liège, Gembloux, Belgium, [2] Evolutionary Biology and Ecology Unit, Faculté des Sciences, Université Libre de Bruxelles, Brussels, Belgium, [3] INRAE, BFP, University of Bordeaux, Villenave d'Ornon, France, [4] INRAE, BIOGECO, University of Bordeaux, Cestas, France, [5] Department of Geography, College of Life and Environmental Sciences, University of Exeter, Exeter, United Kingdom, [6] Evolutionary Genomics, Centre for Geogenetics – Natural History Museum of Denmark, Copenhagen, Denmark

OPEN ACCESS

Edited by:
Juan Viruel, Royal Botanic Gardens, Kew, United Kingdom

Reviewed by:
Carolina Carrizo García, Instituto Multidisciplinario de Biologia Vegetal (IMBIV), Argentina
Jun Ying Lim, Nanyang Technological University, Singapore

***Correspondence:**
Armel S. L. Donkpegan
armel.donkpegan@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: *13 December 2019*
Accepted: *19 May 2020*
Published: *17 June 2020*

Citation:
Donkpegan ASL, Doucet J-L, Hardy OJ, Heuertz M and Piñeiro R (2020) Miocene Diversification in the Savannahs Precedes Tetraploid Rainforest Radiation in the African Tree Genus Afzelia (Detarioideae, Fabaceae). Front. Plant Sci. 11:798. doi: 10.3389/fpls.2020.00798

The dating of diversification events, including transitions between biomes, is key to elucidate the processes that underlie the assembly and evolution of tropical biodiversity. *Afzelia* is a widespread genus of tropical trees, threatened by exploitation for its valuable timber, that presents an interesting system to investigate diversification events in Africa. Africa hosts diploid *Afzelia* species in the savannahs north and south of the Guineo-Congolian rainforest and autotetraploid species confined to the rainforest. Species delimitation and phylogenetic relationships among the diploid and tetraploid species remained unresolved in previous studies using small amounts of DNA sequence data. We used genotyping-by-sequencing in the five widespread *Afzelia* species in Africa, the savannah species *A. africana* and *A. quanzensis* and the rainforest species *A. bipindensis*, *A. pachyloba,* and *A. bella*. Maximum likelihood and coalescent approaches resolved all species as monophyletic and placed the savannah and rainforest taxa into two separate clades corresponding to contrasted ploidy levels. Our data are thus compatible with a single biome shift in *Afzelia* in Africa, although we were unable to conclude on its direction. SNAPP calibrated species trees show that the savannah diploids started to diversify early, at 12 (9.09–14.89) Ma, which contrasts with a recent and rapid diversification of the rainforest tetraploid clade, starting at 4.22 (3.12 – 5.36) Ma. This finding of older diversification in a tropical savannah clade vs. its sister rainforest clade is exceptional; it stands in opposition to the predominant observation of young ages for savannahs lineages in tropical regions during the relatively recent expansion of the savannah biome.

Keywords: *Afzelia*, Leguminosae (Detarioideae), high-throughput sequencing, phylogenomics, coalescent approaches, biome shift, molecular dating, species trees

INTRODUCTION

The biogeographic history of the African flora has been marked by an overall trend toward continental aridification since the wet and warm conditions of the Paleocene (66 – 56 Ma, Senut et al., 2009; Jacobs et al., 2010). Paleobotanical evidence from the north of Africa suggests that rainforest was the most common biome during the Paleocene and the beginning of the Eocene (56 – 33.9 Ma, Jacobs, 2004; Jacobs et al., 2010). More open vegetation appeared in central Africa in the middle Eocene (47.8 – 38 Ma) concomitant with increased temperatures and aridification (Jacobs et al., 2010). A global cooling at the Eocene-Oligocene boundary (33 Ma) led to large-scale extinctions (Zachos et al., 2008; Jacobs et al., 2010) and the grass-dominated savannah biome began to expand in the middle Miocene (16 Ma, Jacobs, 2004), becoming a well established component of tropical vegetation from the late Miocene (ca. 8 Ma, Cerling et al., 1997). The alternation of cold/dry and hot/humid climates of the Miocene (23 – 5.3 Ma), Pliocene (5.3 – 2.6 Ma) and Pleistocene (2.6 – 0.01 Ma) has affected the distribution the major tropical biomes - rainforest, woodland and savannah – with repetitive phases of major expansion or contraction, resulting in the modern distribution of tropical African biomes (Sarnthein and Fenner, 1988; Morley, 2000; Plana, 2004; Salzmann and Hoelzmann, 2005; Anhuf et al., 2006; Miller and Gosling, 2014).

These historical contractions and expansions of the major African biomes have probably triggered biome shifts and diversification in the evolution of tropical plant lineages. Understanding and dating biome shifts is key to understanding the processes that underlie the assembly and evolution of African tropical biodiversity (Wiens and Donoghue, 2004), however, biome shifts have been little studied in the African floras. Multiple transitions from rainforests to dry forests/savannahs have been inferred in the diversification of the tree genus *Guibourtia* in Africa (Tosso et al., 2018). Similarly, three biome transitions from humid forest to dry or montane forests have been documented in the tree genus *Entandrophragma* (Meliaceae) along with ecological adaptations to drier habitat (Monthe et al., 2019). The literature suggests that most biome shifts in tropical Africa support the transition from closed habitats to open habitats (Holstein and Renner, 2011; Veranso-Libalah et al., 2018), which is congruent with paleobotanical evidence for rainforest to be ancient and savannahs to be a more recent biome (Jacobs et al., 2010).

In African forest trees, phylogenetics or population genetics studies have led to the discovery of many new species that could not *a priori* be distinguished based on morphological features (Koffi et al., 2010; Duminil et al., 2012; Heuertz et al., 2014; Daïnou et al., 2016; Ikabanga et al., 2017; Lissambou et al., 2018). High-throughput sequencing can facilitate the study of taxonomically difficult groups that contain closely related, weakly differentiated species. Sequencing large portions of the genome of non-model organisms can help generate resolved phylogenies of these complex groups. For non-model taxa, reduced representation sequencing methods such as genotyping-by-sequencing (GBS, – Elshire et al., 2011) can provide thousands of single nucleotide polymorphisms (SNPs) for phylogenetic analysis without prior knowledge of the genome (Eaton and Ree, 2013; Escudero et al., 2014; Hipp et al., 2014; Ariani et al., 2016; Nicotra et al., 2016; Fernández-Mazuecos et al., 2018).

Afzelia Smith (Detarioideae – Caesalpinioideae) is a widespread and taxonomically complex genus of valuable timber trees that provides an excellent opportunity to apply genomic tools for species delimitation and investigate the role played by biome shifts in species diversification in tropical Africa. *Afzelia* is a Paleotropical genus distributed in Sub-Saharan Africa, where it is known as "doussié," and Southeast Asia (Donkpegan et al., 2014). The genus exhibits large morphological variability within and between species and can be considered a species complex (Donkpegan et al., 2014). At present, most taxonomists agree that it contains 11 species (Chevalier, 1940; Léonard, 1950; Institut National pour l'Etude Agronomique du Congo-belge [INEAC], 1952; Aubréville, 1959, 1968, 1970; Satabié, 1994). Seven species occur in sub-Saharan Africa: five of them are widely distributed, the savannah species *A. africana* Sm. ex Pers., and *A. quanzensis* Welw., and the rainforest species *A. bipindensis* Harms, *A. bella* Harms, *A. pachyloba* Harms (**Figure 1**); and two are local endemics, *A. parviflora* (Vahl) Hepper occurring in rainforest habitat in West Africa, and *A. peturei* De Wild, probably the least documented species of the *Afzelia* clade in Africa, being found in the transition zone between the rainforest and the Zambesian savannah. The remaining four species, *A. xylocarpa* (Kurz) Craib, *A. rhomboidea* (Blanco) S. Vidal, *A. javanica* (Miq) J. Léonard and *A. palembanica* Baker, occur in Southeast Asia in scattered locations in dry, mixed deciduous or evergreen dipterocarp forest. Based on a fossil attributed to *Afzelia* discovered in the Guang River flora in north-western Ethiopia and dating from the Late Oligocene (27.23 Ma, Pan et al., 2010), it is likely that the genus originated in Africa and that it dispersed subsequently into tropical Asia. Most *Afzelia* species are categorized as vulnerable according to the International Union for the Conservation of Nature (IUCN) Red List because they are substantially exploited for the international timber market (International Union for Conservation of Nature and Natural Resources [IUCN], 2012).

In the evolution of *Afzelia* species in Africa, biome shifts seem to have taken place in association with ploidy levels. The rainforest species *A. bipindensis*, *A. bella*, *A. pachyloba*, and *A. parviflora* – sympatrically distributed across the Guineo-Congolian rainforest – have recently been shown to be autotetraploids using nuclear microsatellites and flow cytometry, whereas the savannah species *A. africana* and *A. quanzensis* – situated north and south of the Guineo-Congolian rainforest, respectively – are diploids (Donkpegan et al., 2015). However, based on Sanger sequencing of two nuclear (nDNA) and three plastid (pDNA) regions, and on full plastome sequences for each of the species, the phylogenetic relationships in the polyploid complex remain uncertain. First, the phylogenetic relationships between the forest and the savannah species were not resolved. Therefore, it remains uncertain whether the forest-savannah transitions have happened once or multiple times in the evolution of the genus. There was cyto-nuclear incongruence with respect to the placement of *A. quanzensis*: pDNA placed the savannah species *A. quanzensis* as sister to the tetraploid forest

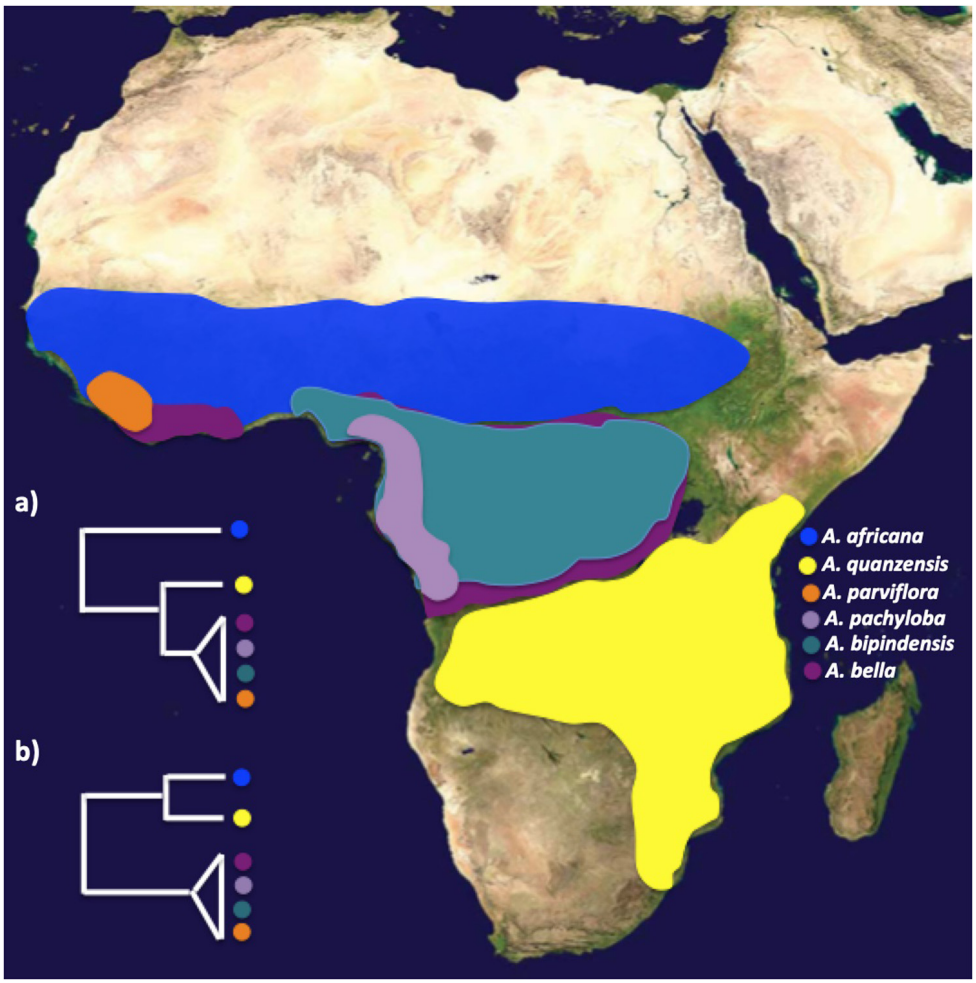

FIGURE 1 | Biogeographic ranges of *Afzelia* sequenced in this study, and alternative phylogenetic relationships **(a,b)** recovered from a previous study (Donkpegan et al., 2017). Plastid data (*psbA*, *trnL*, *ndhF* and genome-wide SNPs via plastomes) suggest that savannah species are paraphyletic with respect to forest taxa **(a)**, whereas nuclear markers (ribosomal *ITS* and the single-copy *PEPC* E7 gene) recovered distinct savannah and tropical forest clades **(b)**. Map image: public domain from www.simplemappr.net.

clade, whereas nuclear markers identified the savannah and the forest species as two monophyletic sister clades (**Figures 1a,b**; Donkpegan et al., 2017). Second, the forest taxa showed little genetic differentiation and displayed extensive plastid and nDNA haplotype sharing across species based on few genetic markers, which could be due to incorrect taxonomy, recent speciation with large effective population sizes or ongoing hybridization (Pennington and Lavin, 2016), whereas the savannah species were genetically well differentiated (Donkpegan et al., 2017, 2020). A phylogeny based on genome-wide genetic markers has the potential to improve the taxonomic classification and our understanding of the evolutionary history of this genus of economically important African trees, including the history of biome shifts between forests and savannahs, and shed light on the speciation process in the rainforest taxa.

In this study, we used GBS to sequence the five most abundant species of the genus *Afzelia* in Africa (*A. africana*, *A. quanzensis*, *A. bipindensis*, *A. bella,* and *A. pachyloba*) in order to assess the phylogenetic relationships among them using multiple methods and datasets. We addressed the following questions.

1. Given the previously unresolved phylogeny, can genome-wide genetic markers provide additional insights into the phylogenetic relationships between diploid savannah and tetraploid rainforest species in *Afzelia*?
2. If so, can molecular dating of the phylogeny inform on which biome shifts occurred during the diversification of *Afzelia* in Africa?
3. Given the extensive haplotype sharing previously observed, can multiple genomic markers delimit species and provide insights into the timing of diversification and/or hybridization in the polyploid complex of rainforest *Afzelia* taxa?

We find strong support for the delimitation of the investigated *Afzelia* species and for phylogenetic relationships between

species. This study represents the most comprehensive phylogenomic evaluation of *Afzelia* to date.

MATERIALS AND METHODS

Sampling, DNA Extraction, Genomic Libraries and Sequencing

We used 41 accessions of *Afzelia* and six accessions (**Supplementary Material S1**) of other Leguminosae species as outgroups. Our sampling represents the five widely distributed species of the genus *Afzelia* in Africa (Donkpegan, 2017): the diploid savannah species *A. africana* (12 accessions) and *A. quanzensis* (7 accessions) and the tetraploid rainforest species *A. bipindensis* (14 accessions), *A. bella* (4 accessions) *A. pachyloba* Harms (4 accessions). Two species of *Afzelia* (*A. parviflora* and *A. peturei*), which have very restricted ranges, were not included in this study due to lack of recently collected plant material required for GBS. The outgroups were chosen according to the latest available phylogenies in the legume family (Bruneau et al., 2008; LPWG, 2017): namely *Scorodophloeus zenkeri* Harms (one accession), *Prioria balsamifera* (Vermoesen) Breteler (2 accessions), *Prioria oxyphylla* (Harms) Breteler (2 accessions), *Peltogyne sp*. (one accession) and the putative sister species to the *Afzelia* clade, *Intsia bijuga* (Colebr.) Kuntze (one accession). Metadata on all accessions are given in **Supplementary Material S1**.

DNA was extracted from silica-dried leaves collected in the field and four recent herbarium specimens (National Herbarium of the Netherlands Wageningen, WAG; African Botanical Library of Université Libre de Bruxelles, BRLU; and the Botanic Garden Meise, BR). For each accession, total genomic DNA was extracted using a CTAB protocol (Doyle and Doyle, 1987) and further purified using the QIAquick method (Qiagen, Venlo, Netherlands). We then quantified and controlled the quality of DNA using a QIAxcel (Qiagen). Before library construction, DNA extracts were further purified using a ZR-96 DNA Clean up kit (Zymo Research, Orange, CA, United States) to remove secondary metabolites. DNA quality was checked on a 1.5% agarose gel and DNA quantity was measured with Qbit HS (Thermo Fisher Scientific, Karlsruhe, Germany). To reach the high concentrations required for the genotyping-by-sequencing (GBS) protocol, two extractions per individual sample were pooled at this second purification step whenever possible.

Overall, 180 GBS libraries were built and sequenced on two Illumina lanes (HiSeq2000, San Diego, CA, United States), using 100-bp Single Read chemistry. Given the large genome sizes of our study species (4.9 – 5.0 pg in the diploids and 8.5 – 9.9 in the tetraploids, Donkpegan et al., 2015), two or three independent GBS libraries per individual were built and sequenced for the diploid and tetraploid individuals, respectively. GBS was performed at the Institute for Genomic Diversity and Computational Biology Service Unit at Cornell University (Ithaca, NY, United States) according to a published protocol (Elshire et al., 2011). To select the best enzyme for the GBS protocol, one microgram of DNA of *Afzelia bipindensis* was used to build test libraries using three different enzymes: *Ape*KI (4.5-base cutter), *Eco*T22I and *Pst*I (both 6-base cutters). Libraries were checked for appropriate fragment sizes (<500 bp) and distribution on an Experion automated electrophoresis system (Bio-Rad Laboratories, Hercules, CA, United States). The enzyme *Eco*T22I gave appropriate fragment sizes (<500 bp) and was selected.

Bioinformatics Analyses

De novo Assembly of Reference Sequence

Single-end reads were checked for quality using FastQC 0.11.5 software (Andrews, 2010). Fastq-formatted GBS data was demultiplexed with Saber software[1]. Low quality bases and adapter contamination were removed with TRIMMOMATIC version 0.33 (Bolger et al., 2014) with the following options: ILLUMINACLIP 2:30:10, LEADING 3, TRAILING 3, SLIDINGWINDOW 4:15, MINLEN 36. The trimmed reads of all *Afzelia* accessions were *de novo* assembled using PyRAD v.3.0.2 software (Eaton and Ree, 2013, see parameters file in **Supplementary Material S2**): sequences were clustered within individuals using VSEARCH allowing for indels and nucleotide polymorphisms and assuming a minimum similarity rate of 85% (Rognes et al., 2016). Consensus allele sequences of each cluster (GBS locus) were generated for each individual based on the jointly estimated heterozygosity (H) and the error rate (E). A two-step approach was used to enrich the final dataset in *Afzelia* loci, and genotype all accessions including outgroups for the same loci. In a first step, clustering was performed using all *Afzelia* accessions in order to build a reference catalog with a minimum similarity rate of 85%. Preliminary genotyping was conducted with PyRAD assuming a minimum depth of 8 and a maximum of four shared polymorphic sites across individuals, to minimize the inclusion of paralogs. In a second step, the best *Afzelia* accession was selected, based on the lowest amount of missing data, as the reference catalog of GBS loci in PyRAD for read mapping and final SNP calling (next section) of all accessions, including outgroups.

SNP Discovery and Genotyping

The trimmed reads of all accessions, including outgroups, were then aligned to the reference sequence using the Burrows-Wheeler Aligner BWA mem 0.7.5a-r405 (Li and Durbin, 2009) with -M and -B 4 options, to generate SAM files. SAM files were processed using SAMtools 0.1.17 (Li et al., 2009) and Picard Tools v1.96[2] to convert from SAM to BAM (Binary Alignment Map) (Sam Format Converter module), sorting the BAM files by position (Sort Sam module) and adding read groups (AddOrReplaceReadGroups module). The resulting BAM files were used as input for Genome Analysis Toolkit (GATK) v3.7 (Depristo et al., 2011). HaplotypeCaller variant discovery was run using emitRefConfidence GVCF mode separately for each sample (Depristo et al., 2011) and GenotypeGVCFs was run on the combined GVCF files to call

[1]https://github.com/najoshi/sabre

[2]http://sourceforge.net/projects/picard/files/picard-tools/1.96/

and genotype SNPs. Although both diploids and tetraploids were present in the dataset, we used a diploid genotyping model in GATK to facilitate analyses involving both ploidy levels and because this practice is known to provide robust results, with only minor loss in numbers of SNPs in the tetraploids (Anderson et al., 2017). VCFtoolsv0.1.15 (Danecek et al., 2011) was used to filter out indels and non-biallelic variants. Different filter criteria were tested for missing data per sample and per SNP (see section "Results"). To evaluate the utility of GBS to infer phylogenetic relationships in *Afzelia*, we considered two separate datasets: the first containing all *Afzelia* samples with the outgroup taxa (hereafter *Afzelia with outgroups* dataset) and the second containing *Afzelia* samples only (hereafter *Afzelia dataset*). The filtered VCF files were converted to a fasta files containing a single consensus sequence per individual using PGDSpider version 2.1.1.5 (Lischer and Excoffier, 2012).

Phylogenetic Analyses and Estimation of Divergence Times

For the *Afzelia with outgroups* dataset we applied Maximum Likelihood (ML) methods to perform phylogenetic analyses. ML analyses were conducted with the GTR + GAMMA substitution model and 100 bootstrap replicates running RAxML 7.2.6 (Stamatakis, 2014) with default parameters through the CIPRES Portal 2.1 (Miller, 2009)[3]. Phylogenetic trees were visualized in FigTree 1.4.3 (Rambaut, 2007). The analyses were repeated with ascertainment bias correction using the Lewis method to avoid overestimation of branch lengths and biases in the phylogeny when the number of non-variable sites is not known (Lewis, 2001). We then estimated divergence times within *Afzelia* based on the resulting phylogeny using Bayesian MCMC analysis implemented in BEAST 1.7.4 (Drummond et al., 2012). To facilitate comparison with previously estimated divergence times, we used an *Afzelia* fossil dated to 27.23 Ma (Pan et al., 2010) as in Donkpegan et al. (2017). Since our RAxML phylogeny was unable to resolve the positioning of *Intsia* with respect to *Afzelia* and because of the similarity of species from both genera for morphological characters used to describe the fossil (Kadiri and Olowokudejo, 2008; Pan et al., 2010), we considered the fossil to represent the minimum age for diversification of the *Afzelia-Intsia* clade. Bayesian analyses were done using the following priors: an uncorrelated lognormal relaxed clock model, a Yule process of speciation which is adequate to analyze data at the interspecific level (Yule, 1925; Heled and Drummond, 2015), and the selected nucleotide substitution model. The MCMC analyses were run for 50,000,000 generations, sampling trees every 1,000 generations. To evaluate convergence and ensure sufficient effective sample sizes (ESS values) for all BEAST parameters, we used Tracer 1.6 (Rambaut and Drummond, 2016). Runs were combined with LogCombiner after removing the first 10,000 samples as burn-in. Maximum Clade Credibility trees were produced in TreeAnnotator 1.8 (Drummond and Rambaut, 2007) and plotted in FigTree 1.4.4[4].

Species Tree Inference: Species Delimitation and Estimation of Divergence Times

To provide additional support for species delimitation in *Afzelia*, we used several models to estimate species trees directly from the *Afzelia dataset*. We used the multi-species coalescent approach in SNAPP v.1.3.0 (Bryant et al., 2012) implemented in BEAST2 v.2.5.2 (Bouckaert et al., 2014) to estimate species trees from a multilocus SNP matrix. SNAPP is based on a Bayes Factor Delimitation method (BFD*) (Leaché et al., 2015) which allows for the comparison of alternative species delimitation models in an explicit multispecies coalescent framework. The corresponding VCF file was converted to PHYLIP format using the Python script "vcf2phylip"[5] and the XML input file for SNAPP analyses was prepared using the Ruby script "snapp_prep.rb"[6]. Two independent Markov-Chain Monte Carlo (MCMC) simulations were run for one million generations each, sampling trees at 1000 step intervals. Stationarity and convergence of chains were visually checked in TRACER v1.6 (ESS > 1000; Rambaut and Drummond, 2016)[7]. The program DensiTree v.2.2.6 (Bouckaert and Heled, 2014) was used to visualize the SNAPP trees after discarding the first 10% of each MCMC chain as burn-in. The resulting tree and log files were combined with Logcombiner 1.8.2[8] with a burn-in of 10 000 for each run. A maximum-clade-credibility summary tree was generated with TreeAnnotator (Drummond and Rambaut, 2007) and visualized in FigTree v.1.4.3[9]. Divergence dating based on the multispecies coalescent has been shown to provide more accurate results than dating of concatenation-based species trees for the ages of younger nodes, which are commonly overestimated when concatenation is used (Stange et al., 2018). We dated the resulting species tree in SNAPP using the same fossil as above. To generate a SNAPP input XML file for divergence dating, we followed the protocol of Stange et al. (2018), using their provided Ruby script ("snapp_prep.rb"). According to their approach, a molecular clock and effective population sizes are shared between all species. We assumed the age of the root to be within a normal distribution of mean = 27.23 Ma and standard deviation being 10% of that variation, with an offset of 20 Ma.

We then used the Generalized Mixed Yule Coalescent (GMYC) model (Pons et al., 2006; Fujisawa and Barraclough, 2013) based on a likelihood method for species delimitation in *Afzelia*. To account for uncertainty in species delimitation, we used the single (*sGMYC*) and multiple (*mGMYC*) threshold species delimitation models using packages APE (Paradis et al., 2004) and SPLITS (Ezard et al., 2013) in R (R Core Team, 2017). The GMYC method requires an ultrametric

[3]www.phylo.org

[4]http://tree.bio.ed.ac.uk/software/figtree/

[5]https://github.com/edgardomortiz/vcf2phylip

[6]https://github.com/mmatschiner/snapp_prep

[7]http://beast.bio.ed.ac.uk/tracer/

[8]http://beast.bio.ed.ac.uk/LogCombiner

[9]http://tree.bio.ed.ac.uk/software/figtree/

tree (i.e., calibrated with a molecular clock), which was constructed using BEAST v.1.8 (Bouckaert et al., 2014). We used a relaxed log-normal clock with a coalescent tree prior as these have been identified as best prior parameters for GMYC analyses (Esselstyn et al., 2012). Monte Carlo Markov chains (MCMC) were run for one million iterations, sampling every 1000 iterations. Convergence of chains was assessed using Tracer v.1.6 (Rambaut and Drummond, 2016). The consensus tree (maximum clade credibility tree; 10% burn in; tree not presented) was constructed with TreeAnnotator v.1.7 (Drummond and Rambaut, 2007).

Finally, we used Bayesian (*bPTP*) and ML (*mlPTP*) implementations of the Poisson tree processes model (PTP) (available at http://species.h-its.org/ptp/) to estimate the number of speciation events in the *Afzelia* rooted phylogenetic tree based on nucleotide substitutions (Zhang et al., 2013). Because this approach does not require ultrametrization of trees, it constitutes a reasonable alternative to other species delimitation models such as the General mixed Yule coalescent model (Pons et al., 2006). In PTP models, the numbers of substitutions (branch lengths) represent speciation or branching events and, therefore, they only require a phylogenetic tree as input. PTP analyses were conducted on the web server for PTP (available at http://species.h-its.org/ptp/) using the best ML tree resulting from the RA × ML analysis (see below).

Inference of Interspecific Hybridization History

We used TreeMix v1.12 (Pickrell and Pritchard, 2012) to infer historical relationships among *Afzelia* species. This method builds a maximum likelihood graph that connects species with their common ancestor, using the covariance structure of allele frequencies between species and a Gaussian approximation for genetic drift. Migration events, i.e., hybridization events, can be modeled to improve the fit of the inferred graph. To meet TreeMix requirements, the *Afzelia dataset* was reduced to SNPs without missing data, a single SNP was selected per GBS locus using VCFtools, and species-level allele frequencies were computed from the VCF to generate the TreeMix infile. We modeled the interspecific evolutionary history in *Afzelia* using $m = 0$ to $m = 4$ migration events.

RESULTS

GBS Data Production and Reference Sequence Construction

Unambiguous barcodes were found in a total of 295 million sequencing reads. After trimming, cleaning, and quality filtering an average of 4.7 million reads per accession were retained. Genotyping *Afzelia* accessions with PyRAD yielded accession-level heterozygosity estimates between 0.0288 and 0.0794 and error rates between 0.0043 and 0.0169 (**Supplementary Material S3**). The accession with the lowest amount of missing data in the PyRAD genotyping – AD657 of *A. bipindensis* – was used as a GBS reference sequence. It comprised a total of 221,334 loci representing 3,489,577 bp, including 52,314 polymorphic sites, and 3749 scaffolds (**Supplementary Material S3**). This reference is available in FASTA format in DRYAD[10].

Mapping and SNP Calling

Mapping the reads from all accessions against the reference and genotyping with GATK allowed us to obtain VCF files for two datasets. For the *Afzelia with outgroups* dataset 21,150 SNPs were discovered and 9,165 SNPs were retained in 26 *Afzelia* accessions and all seven outgroups, after filtering INDELs, non-biallelic sites, and sites with more than 40% of missing data. For the *Afzelia* dataset 23,694 SNPs were discovered and 4,823 SNPs were retained (26 accessions) after filtering INDELs, non-biallelic sites, and sites with more than 20% of missing data. The *Afzelia with outgroups* and *Afzelia* datasets were used to generate phylogenetic trees in RAxML. For species delimitation based on the multispecies coalescent in SNAPP, a subset of 2,370 bi-allelic SNPs without missing data in at least one species was retained.

Concatenation-Based Tree and Timing of *Afzelia* Diversification

For the *Afzelia with outgroups* dataset, different datasets (on the percentage of missing data) were tested for phylogenomics of the genus. The dataset (9,165 SNPs) of at most 40% missing data yielded the phylogenetic relationships that were most congruent with the known topology in the legumes (Bruneau et al., 2008; LPWG, 2017). The *Afzelia* clade (including *Intsia bijuga*) formed a monophyletic group in the RAxML tree (**Figure 2**). We also obtained two strongly supported monophyletic clades that correspond to the diploid and tetraploid lineages in the *Afzelia-Intsia* clade. All *Afzelia* species – represented by multiple accessions each – appeared monophyletic with strong support for four species and lower support for *A. bipindensis*. Such well-supported species delimitation had never been found in previous studies of phylogeny in the genus, especially in the rainforest clade of tetraploid species. The differentiation of the savannah species was well supported. In contrast, the topology of the diversification of the three rainforest species was not well resolved, as revealed by low bootstrap supports. The analyses with ascertainment bias correction to avoid overestimation of branch lengths and biases in the phylogeny using SNP datasets, resulted in identical topology and did not improve the resolution of the rainforest clades (results not shown).

Based on molecular dating of the concatenation-based phylogeny, diversification appeared to occur earliest in the diploid lineage of *Afzelia*, while the position of the sister species *Intsia bijuga* remained ambiguous on the topology of the tree (**Figure 3**). The diversification of the *Afzelia-Intsia* clade started in the Oligocene, the posterior mean age of the common ancestor (MRCA, node A) of the clade being estimated at 33.31 Ma [95% highest posterior density (HPD) 28.64–41.04 Ma] (**Figure 3**). The divergence of each monophyletic species on the basis of the well-resolved phylogenetic tree at the genus level suggests that savannah species diversified in the Middle Miocene, *A. africana* (node F, 12.33 Ma) and *A. quanzensis* (node E, 14.68 Ma),

[10]https://doi.org/10.5061/dryad.95~x~69p8gf

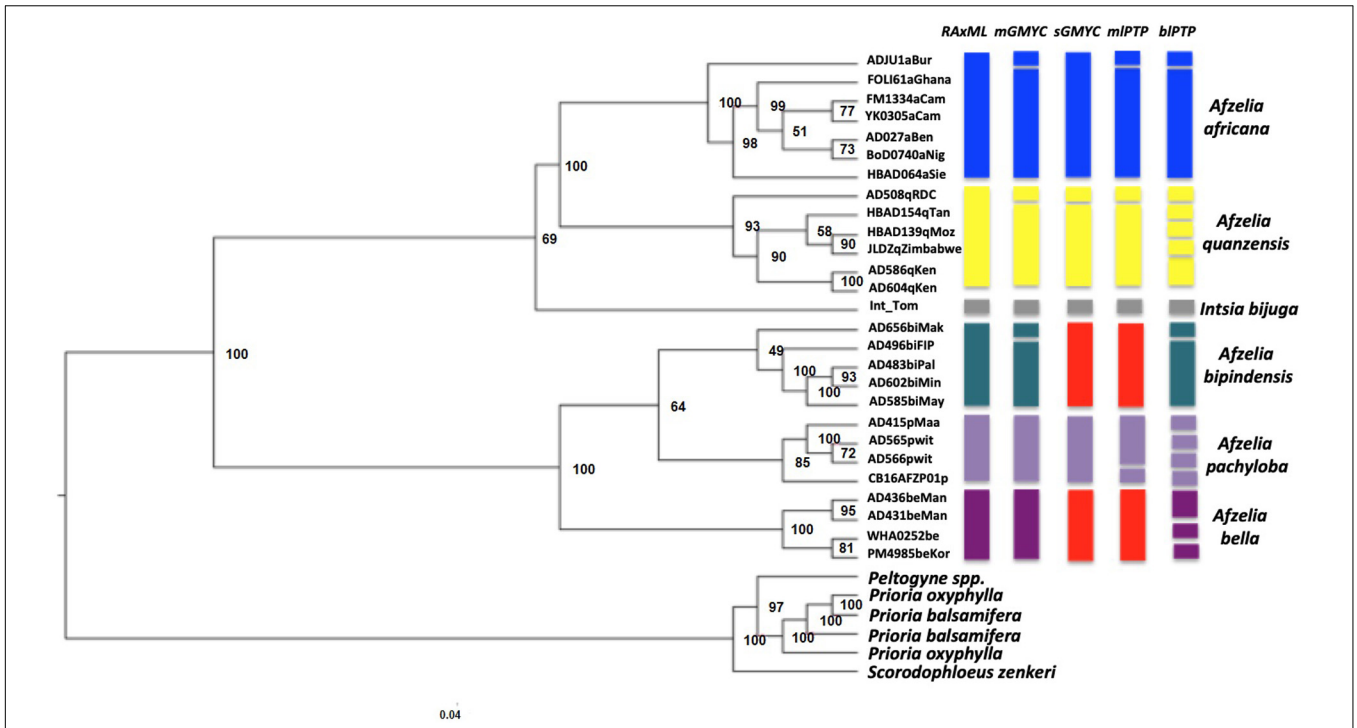

FIGURE 2 | Phylogenetic relationships in *Afzelia* and related taxa inferred from nuclear genomic data. Maximum likelihood tree (33 samples, 9165 SNPs) estimated in RaxML. The tree was created using a 50% majority rule consensus tree from 500 bootstrap replicates. The consensus multilocus coalescent species trees of the genus *Afzelia* based on four models are represented in the right part of the figure using colored blocks. Each botanically determined species is indicated with the same branch color as in **Figure 1**. Each separate block stands for a separate lineage or taxonomic entity as delimited with the species delimitation model noted above. The red color representing *A. bipindensis* and *A. bella* in sGMYC and mlPTP indicates that the two species shared the same clade for these models.

whereas the rainforest species would have appeared in the Upper Miocene: *A. bipindensis* (node G, 10.06 Ma), *A. pachyloba* (node H, 10.08 Ma) and *A. bella* (node I, 08.39 Ma).

Coalescent-Based Species Trees and Timing of *Afzelia* Diversification

For the *Afzelia* dataset the SNAPP analysis resolved five well-differentiated clades that support the monophyly of all species. The visualization of the species trees superimposed in the DensiTree plot (**Figure 4**) resolved the evolutionary relationships among species with no signs of conflict among trees, even within the rainforest clade: *A. pachyloba* is the sister species of the clade containing *A. bipindensis* and *A. bella*. The results of the *GMYC* and *PTP* models are plotted against the RAxML phylogeny in **Figure 2**. They resolved between 6 and 15 sub-lineages within *Afzelia*. The *sGMYC* and *mlPTP* models placed *A. bella* and *A. bipindensis* into the same species cluster, in line with the close relationship revealed by the SNAPP species tree. The SNAPP-calibrated tree revealed, as expected, later diversification dates than the concatenated gene tree. The diversification of *Afzelia* has a the posterior mean age of the MRCA in the late Oligocene at 26.93 Ma (95% HPD 21.06 – 32.92 Ma; **Figure 4**). The rainforest lineage diversified rapidly between the Pliocene and the early Pleistocene (mean *A. pachyloba* split at 4.22 Ma and mean *A. bipindensis*/*A. bella* split at 2.78 Ma) and the savannah lineage earlier in the Miocene (mean at 12 Ma).

Interspecific Hybridization History

The genetic relationships among species revealed by TreeMix distinguished the diploid and tetraploid clades, in agreement with the phylogenetic relationships evidenced with concatenation or coalescent-based methods, and confirmed the placement of *A. pachyloba* as sister of the clade containing *A. bipindensis* and *A. bella*, as revealed by species delimitation methods (**Figure 5A**). The proportion of variance in the data explained by the model was high, PVE = 0.975, for a model without migration. The addition of migration events improved the proportion of variance explained to PVE = 0.998 for m = 1 and PVE = 0.9999 for m = 2 migration events. The first migration event links diploid *A. africana* with tetraploid *A. bella* whereas the second links an ancestor of diploid *A. quanzensis* with tetraploid *A. bipindensis* (**Figure 5**).

DISCUSSION

Our phylogenetic reconstructions provided the most robust phylogenetic framework of the tropical tree genus *Afzelia* in Africa produced to date. Both the SNP concatenated gene tree (RAxML tree; **Figure 2**) and the coalescent-based species tree (SNAPP tree; **Figure 4**) highly supported two major monophyletic clades associated with habitat and ploidy levels: a diploid savannah clade and a tetraploid

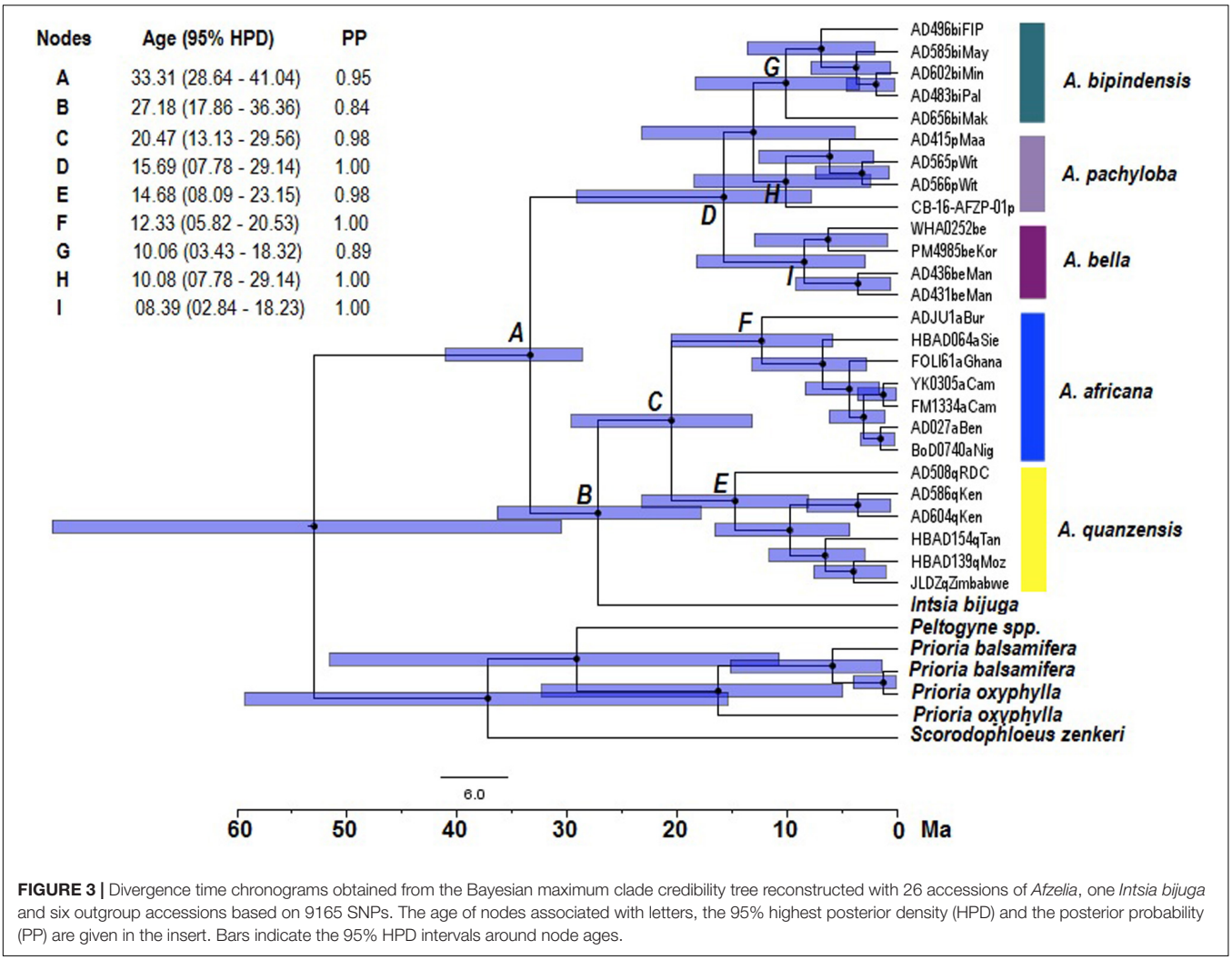

FIGURE 3 | Divergence time chronograms obtained from the Bayesian maximum clade credibility tree reconstructed with 26 accessions of *Afzelia*, one *Intsia bijuga* and six outgroup accessions based on 9165 SNPs. The age of nodes associated with letters, the 95% highest posterior density (HPD) and the posterior probability (PP) are given in the insert. Bars indicate the 95% HPD intervals around node ages.

rainforest clade. The calibrated phylogeny and the species tree (**Figures 3**, **4**) show an earlier diversification of the savannah clade followed by a later speciation within the rainforest clade. Species delimitation within these two major clades, with all species resolved as monophyletic, was also the most robust to date.

A Single Biome Shift in African *Afzelia*

Genotyping-by-sequencing data strongly supported the monophyly of two major habitat-specific clades in *Afzelia*. This suggests a single transition between the savannah and the rainforest biomes in Africa. Previous plastid DNA sequence data – Sanger sequences of a few loci and full plastome sequences – placed the savannah species *A. quanzensis* as sister to the rainforest clade (Donkpegan et al., 2017). In contrast, the phylogenetic analyses of the multilocus genomic GBS data strongly supported the monophyly of the savannah clade and suggested a unique forest-savannah shift in the diversification of *Afzelia* in Africa.

The calibrated phylogeny and the dated species tree indicate a diversification of the savannah clade in the Miocene, earlier than the rainforest clade. This early diversification of the savannah species is exceptional in the context of the global plant evolutionary patterns reported in tropical Africa and South America that point to a relatively young age for savannah lineages. In the South African tree flora, the majority of divergence times between sister taxa of savannah trees were dated within the Pleistocene (the last 2 Ma), which is more recent than those between sister taxa of forest trees (Maurin et al., 2014). In the Brazilian savannah (Cerrado) the origin of woody plants restricted to the Cerrado is also estimated to be recent (<10 Ma), most of them in the Pliocene (<4 Ma; Simon et al., 2009). This seems to reflect the relative ages of the biomes, with rainforests dating from at least the early Paleocene (Burnham and Johnson, 2004; Jacobs et al., 2010), whereas savannahs are much younger, arising only in the Miocene (Jacobs, 2004; Senut et al., 2009; Simon et al., 2009; Pennington and Hughes, 2014).

Evolutionary shifts of plant lineages from the rainforest to savannah or to other dry biomes seem to have been significantly more frequent than switches of plant lineages

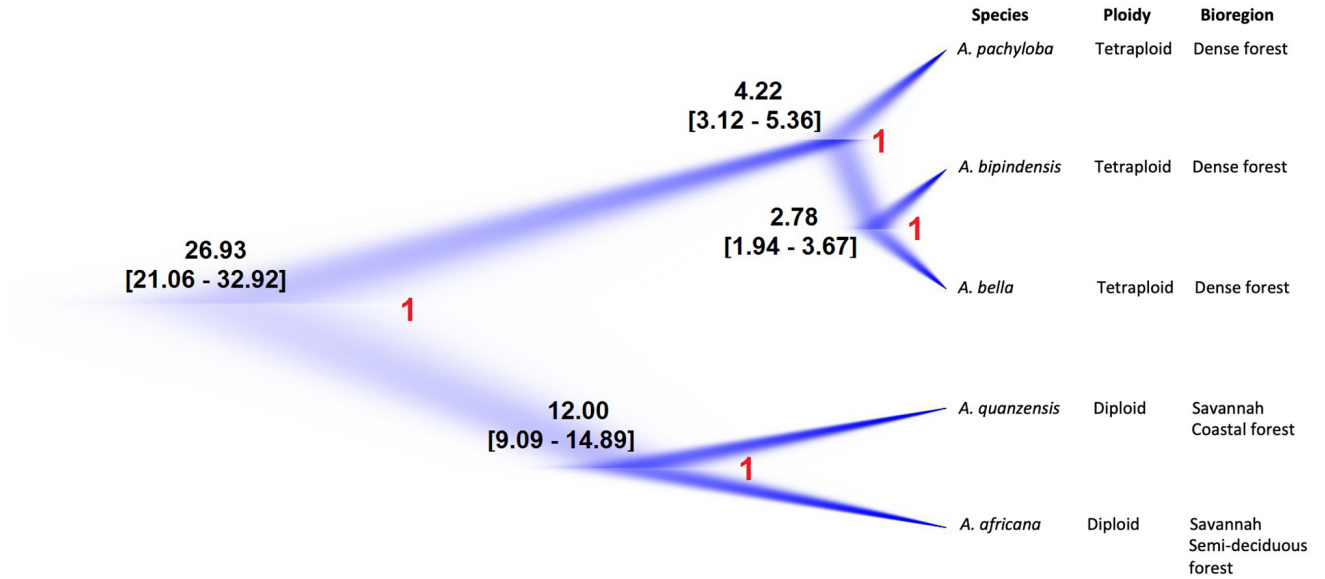

FIGURE 4 | Nuclear species tree estimation in *Afzelia* based on 2370 SNPs as inferred by SNAPP and summary of species' ecological and ploidy characteristics. Maximum-clade-credibility tree is shown in dark blue. Estimated ages have been indicated at each node with 95% highest posterior density (HPD) and posterior probability (in red).

into the rainforest (Simon et al., 2009; Simon and Pennington, 2012, Donoghue and Edwards, 2014, Maurin et al., 2014, Tosso et al., 2018; Freitas et al., 2019; Monthe et al., 2019). It seems logical that older biomes, i.e., the rainforests, will act as sources of lineages for younger biomes, i.e., the savannahs. Of course, there is no *a priori* reason to expect biome switches to be unidirectional, and given enough time, savannah lineages may re-enter rainforests. This might have been the case of the African *Afzelia,* where molecular phylogenies indicate recent splits within the rainforest lineage in the Upper Miocene and Pleistocene. If we assume that the splits of extant species occurring in the same biome indicate a shared ancestor within that biome, it is possible that the African *Afzelia* originated as a diploid savannah lineage followed by whole genome duplication and subsequent diversification in the rainforest. However, we should note that biome conservatism is more common than biome shifts in phylogenies of Southern Hemisphere plants (Crisp et al., 2009) and that the topology of the *Afzelia* phylogeny alone did not allow us to conclude on the direction of the biome shift, as both monophyletic clades diversified into their respective rainforest or savannah biomes.

In terms of morphological trait variation in *Afzelia*, dry forest species (*A. africana* and *A. quanzensis*) are clearly differentiated from rainforests species (*A. bella*, *A. bipindensis*, and *A. parviflora*) based on vegetative and floral variables (Donkpegan et al., 2017). The main morphological discriminant variables which separate dry forest versus rainforest species are (i) the rounded versus truncated-attenuated basal leaflet shape; (ii) at the distal end of the leaflets, the scalloped versus acuminate and mucronate shapes (sometimes with black spots). Leaves are longer and have a smaller number of secondary veins in dry forest vs. rainforest *Afzelia* species (Donkpegan et al., 2017). It is, however, difficult to associate these leaf and vein trait differences with a clear habitat-adaptive role in the light of the literature (Sack and Scoffoni, 2013; Tng et al., 2013) and to our knowledge there are no studies available that treat functional or physiological trait variation in *Afzelia*.

Evolutionary Radiation of the African Rainforest Species

Previous Sanger sequencing of two nuclear and three plastid regions revealed extensive allele sharing across the rainforest clade of *Afzelia*. Accessions of the same species were scattered across the trees (Donkpegan et al., 2017), which could be due to incorrect taxonomy, recent speciation or ongoing hybridization. The GBS data effectively resolved species delimitation and diversification times of *Afzelia* in the rainforest. All species – which are morphologically similar and broadly sympatric- were resolved as monophyletic, validating the latest taxonomic revisions. The data also revealed a recent diversification process, and no trace of hybridization between species.

In *Afzelia*, autopolyploidization occurred prior to rapid speciation in the rainforests, suggesting the role of whole-genome duplications in the onset of adaptive radiations. Polyploidy represents an immediate source of genetic novelty that may promote evolutionary changes and divergence (Wood et al., 2007). Rapid speciation events immediately after polyploidization are well known in the evolution of

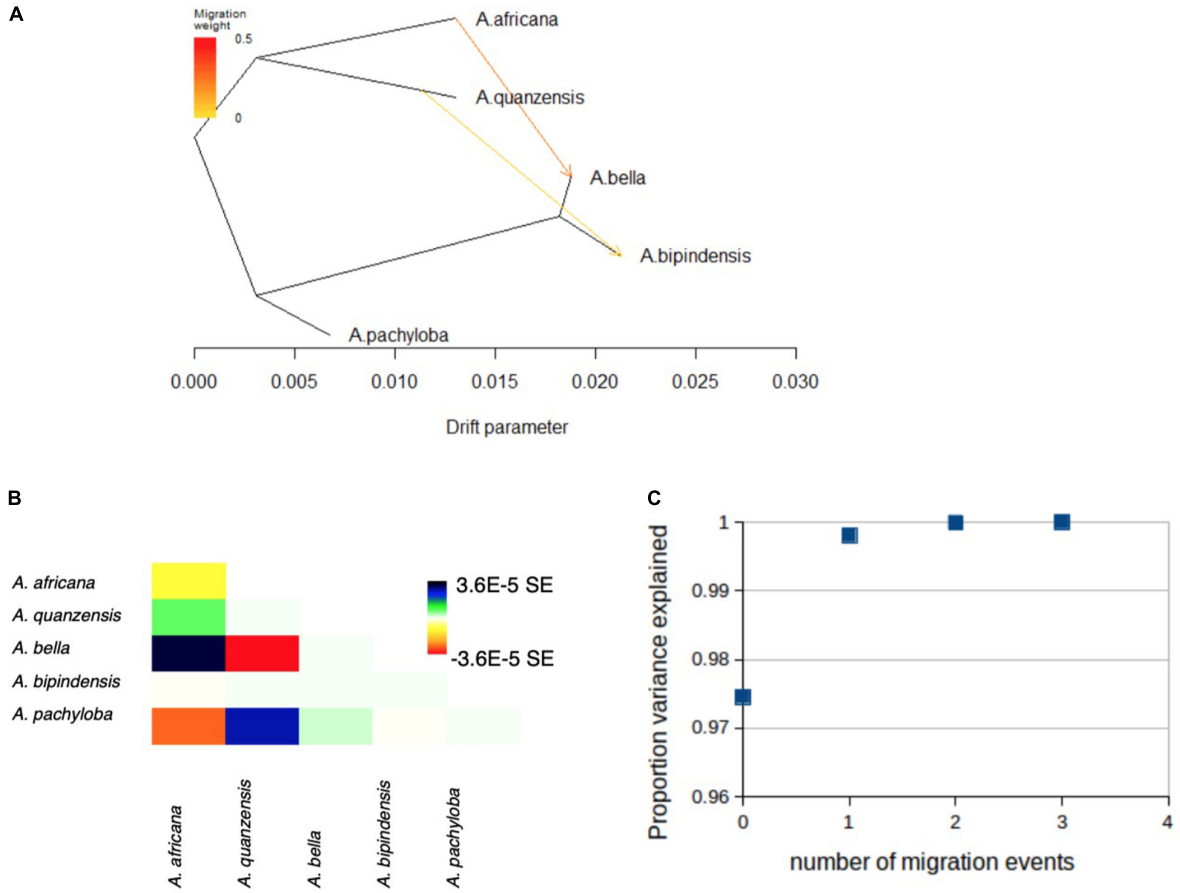

FIGURE 5 | Evolutionary history among *Afzelia* species as inferred by TreeMix. **(A)** Graph showing the topology and branch lengths according to drift parameter, allowing for $m = 2$ migration events, represented by arrows. **(B)** Residual fit for the graph shown in **(A)**. The residual covariance between each pair of species scaled by the average standard error across all pairs is plotted. Colors are described in the palette on the right. Residuals above zero (green and blue) represent species that are more closely related to each other in the data than in the best-fit tree, and thus are candidates for admixture events. **(C)** Proportion of variance of the data explained by the four models run in TreeMix using $m = 0$ to $m = 4$ migration events.

plant groups and point at the success of whole-genome duplications as triggers of speciation (Soltis et al., 2007). The rainforest *Afzelia* species correspond to monophyletic clades, although the clade support varied depending on the filtering parameters used to generate the datasets as well as on the phylogenetic method chosen. This weaker phylogenetic support suggests incomplete lineage sorting in the tetraploid species, which is consistent with their more recent diversification and the larger effective population sizes of tetraploid organisms (Arnold et al., 2012). In addition, the stronger support for monophyly in the savannah than rainforest species is in agreement with similar observations in South America (Pennington and Lavin, 2016). The pattern of strong support for monophyletic species in seasonally dry tropical forests in South America has been attributed to maintenance of resident lineages adapted to a stable, seasonally dry ecology; conversely lower support for monophyly in rainforest trees was attributed to lesser habitat stability creating opportunities for immigration and speciation from taxa with large effective population sizes extending over large areas (Pennington and Lavin, 2016; Heuertz et al., 2020).

Coalescent-Based Phylogenetics and Population Genetics of Multiple Nuclear Loci for Species Delimitation in Recent Radiations

In the rainforest clade of *Afzelia,* phylogenomic analysis revealed short branches among species, in line with a scenario of recent radiation, and revealed the superior performance of the coalescent-based over the concatenation-based methods. Using concatenation-based trees, the topology changed depending on the filtering parameters, which might be explained by the poor performance of this approach under highly incomplete lineage sorting, typically found in cases of rapid speciation (Fernández-Mazuecos et al., 2018). Coalescent-based species trees are based on tests of alternative hypotheses of species delimitation. In this context, lineages do not need to be resolved as monophyletic in gene trees, which leads to more reliable estimates of the

evolutionary relationships and divergence times. Overall, an approach combining multiple nuclear loci with coalescent-based phylogenetic analyses was revealed as the optimal approach to resolve and date the radiation of the rainforest clade.

Forcing the rainforest radiation into bifurcating phylogenetic trees may be problematic because of the age of the radiation and potential for interspecific gene flow. Therefore, we explored the possibility of reticulate evolution using a population genetics approach implemented in TreeMix. Using genome-wide SNPs we found no evidence for gene flow between rainforest species despite their rapid genetic differentiation and sympatric distribution. Given the limited sample sizes within species it is possible that we have missed introgression events limited to sympatric populations of both species. Nevertheless, our analyses identified possible hybridization events that led to incorporation of genetic variation from savannah to rainforest species after the split of the main lineages. Such hybridization events could potentially explain the previously observed cyto-nuclear incongruence, where *A. quanzensis* grouped with rainforest species for plastid DNA but with *A. africana* for two nuclear regions (Donkpegan et al., 2017). *Afzelia quanzensis* could thus have captured plastid DNA from rainforest species during an ancient hybridization event, or else, incomplete lineage sorting may explain the pattern. Altogether, phylogenetic evidence in *Afzelia* points to the existence of hybridization and introgression between species despite differences in ploidy, as has been observed in other plant lineages, such as orchids of the genus *Epidendrum* (Pinheiro et al., 2010) or grasses of the genus *Spartina* (Ainouche et al., 2009).

Similar to Donkpegan et al. (2017) our study did not permit to solve the phylogenetic position of *Intsia bijuga*, the putative sister clade to *Afzelia*. *Intsia bijuga* was placed within the *Afzelia* clade, confirming the close relatedness of both genera. Multilocus data of Asian *Afzelia* species and increased taxon sampling in *Intsia* with multiple accessions per species may help improve the phylogenetic positioning of *Intsia*.

CONCLUSION

We have elucidated the evolutionary history of the widespread emblematic and threatened African tree species of the genus *Afzelia* using genome-wide multilocus data. While the genus was previously recognized as a species complex (Donkpegan et al., 2015), we showed that based on our set of accessions and markers, all species were resolved as monophyletic, and that diversification in the savannah clade preceded that of the rainforest clade. Phylogenomic studies thus represent promising approaches to clarify evolutionary relationships in taxonomic groups that show a high level of variation.

DATA AVAILABILITY STATEMENT

Fasta reference sequence is available from the Dryad data repository: https://datadryad.org/stash/share/QK6Ay8vq7grepB66aGWp42Ir8PCwjrpSR2IXUqX4rRk. Raw fastq sequences of GBS data are being submitted to GenBank's Sequence Read Archive.

AUTHOR CONTRIBUTIONS

AD, MH, and RP conceived the study. All authors collected the data, performed the analyses, interpreted the results, and contributed to drafting and writing the manuscript.

FUNDING

The authors thank the "Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture (FRIA-FNRS, Belgium)," the Marie Curie FP7-PEOPLE-2012-IEF program (project AGORA awarded to RP), the Fonds de la Recherche Scientifique (F.R.S.-FNRS through grant J.0292.17F), the Belgian Science Policy (project AFRIFORD), and the DynAfFor project (funded by FFEM-AFD) for funding this research. This work has also benefited from an "Investissements d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA: ANR-10-LABX-25-01).

ACKNOWLEDGMENTS

The authors acknowledge a "Patrimoine de l'Université de Liège" and Labex COTE mobility grant provided to ASLD at INRAE; Nature+, Esra Kaymak and Barbara Leal for technical assistance, M. Thomas P. Gilbert for hosting the GBS labwork performed by RP at the University of Copenhagen, Jérôme Chave et Bernadette Grosso for providing Peltogyne specimens (BRIDGE collection) and R. Toby Pennington (Geography, College of Life and Environmental Sciences, University of Exeter – United Kingdom) for contributing to the discussion on tropical biome-shifts.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00798/full#supplementary-material

MATERIAL S1 | Sample origins (fresh and herbarium material) of plant tissue samples used for the phylogenetic analyses of *Afzelia*. "na" means that no collector or voucher names are mentioned on the specimen consulted and studied.

MATERIAL S2 | The parameters used in PyRAD for *de novo* assembly.

MATERIAL S3 | Results from *de novo* assembling of 30 *Afzelia* and *Intsia* accessions. (nloci) is the number of loci identified in the intra-accession clustering; (f1loci) is the number of loci with >N depth coverage; (f2loci) number of loci with >N depth and passed paralog filter; (nsites) number of sites across f loci; (npoly) number of polymorphic sites in nsites; (poly) frequency of polymorphic sites. H and E are heterozygosity and error rate, respectively.

REFERENCES

Ainouche, M. L., Fortune, P. M., Salmon, A., Parisod, C., Grandbastien, M. A., Fukunaga, K., et al. (2009). Hybridization, polyploidy and invasion: lessons from *Spartina* (Poaceae). *Biol. Invasions* 11:1159. doi: 10.1007/s10530-008-9383-2

Anderson, B. M., Thiele, K. R., Krauss, S. L., and Barrett, M. D. (2017). Genotyping-by-sequencing in a species complex of Australian hummock grasses (Triodia): methodological insights and phylogenetic resolution. *PLoS One* 12:e0171053. doi: 10.1371/journal.pone.0171053

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 March, 2017).

Anhuf, D., Ledru, M. P., Behling, H., Da Cruz, F. W. Jr., Cordeiro, R. C., Van der Hammen, T., et al. (2006). Paleo-environmental change in Amazonian and African rainforest during the LGM. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 239, 510–527. doi: 10.1016/j.palaeo.2006.01.017

Ariani, A., Berny, M., Teran, J. C., and Gepts, P. (2016). Genome-wide identification of SNPs and copy number variation in common bean (*Phaseolus vulgaris* L.) using genotyping-by-sequencing (GBS). *Mol. Breed.* 36:87.

Arnold, B., Bomblies, K., and Wakeley, J. (2012). Extending coalescent theory to autotetraploids. *Genetics* 192, 195–204. doi: 10.1534/genetics.112.140582

Aubréville, A. (1959). *La flore forestière De La Côte d'Ivoire*, Vol. I., 2 Edn. Nogent-sur-Marne: Centre Technique Forestier Tropical.

Aubréville, A. (1968). Légumineuses – Césalpinioidées. Flore du Gabon. *Muséum Natl. d'Histoire Nat.* 15, 111–118.

Aubréville, A. (1970). Légumineuses - Césalpinioidées (Leguminosae - Caesalpinioideae). *Flore Cameroun* 9:339.

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

Bouckaert, R., and Heled, J. (2014). DensiTree 2: seeing trees through the forest. *bioRxiv* [Preprint]. doi: 10.1101/012401

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

Bruneau, A., Mercure, M., Lewis, G. P., and Herendeen, P. S. (2008). Phylogenetic patterns and diversification in the caesalpinioid legumes. *Botany* 86, 697–718. doi: 10.1139/b08-058

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29, 1917–1932. doi: 10.1093/molbev/mss086

Burnham, R. J., and Johnson, K. (2004). South American palaeobotany and the origins of neotropical rainforests. *Philos. T. Roy. Soc. B* 359, 1595–1610. doi: 10.1098/rstb.2004.1531

Cerling, T. E., Harris, J. M., McFadden, B. J., Leakey, M. G., Quade, J., Eisenmann, V., et al. (1997). Global vegetation change through the Miocene/Pliocene boundary. *Nature* 389, 153–158. doi: 10.1038/38229

Chevalier, A. (1940). Sur un arbre du Cameroun et du Gabon à bois utilisable (Afzelia pachyloba Harms). *Bot. Appl. Agricult. Trop.* 19, 484–488. doi: 10.3406/jatba.1939.6006

Crisp, M. D., Arroyo, M. T., Cook, L. G., Gandolfo, M. A., Jordan, G. J., McGlone, M. S., et al. (2009). Phylogenetic biome conservatism on a global scale. *Nature* 458, 754–756. doi: 10.1038/nature07764

Daïnou, K., Blanc-Jolivet, C., Degen, B., Kimani, P., Ndiade-Bourobou, D., Donkpegan, A. S. L., et al. (2016). Revealing hidden species diversity in sister species using SNPs and SSRs on a systematically collected sample – A case study in the African tree genus Milicia. *BMC Evol.* 16:259. doi: 10.1186/s12862-016-0831-9

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

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498.

Donkpegan, A. S. L. (2017). *Evolutionary History of Afzelia Smith (Leguminosae - Caesalpinioideae) Complex in Forest and Savannah Ecosystems of Tropical Africa*. 175. PhD thesis, University of Liège – Gembloux Agro-Bio Tech, Belgium.

Donkpegan, A. S. L., Doucet, J.-L., Dainou, K., and Hardy, O. J. (2015). Microsatellite development and flow cytometry in the african tree genus Afzelia (Fabaceae. Caesalpinioideae) reveal a polyploid complex. *Appl. Plant Sci.* 3:1400097. doi: 10.3732/apps.1400097

Donkpegan, A. S. L., Doucet, J.-L., Migliore, J., Duminil, J., Daïnou, K., Rosalia, P., et al. (2017). Evolution in African tropical trees displaying ploidy-habitat association: the genus Afzelia (Leguminosae). *Mol. Phyl. Evol.* 107, 270–281. doi: 10.1016/j.ympev.2016.11.004

Donkpegan, A. S. L., Hardy, O. J., Lejeune, P., Oumorou, M., Dainou, K., and Doucet, J.-L. (2014). Un complexe d'espèces d'Afzelia des forêts africaines d'intérêt économique et écologique (synthèse bibliographique). *Biotechnol. Agron. Soc. Environ.* 18, 233–246.

Donkpegan, A. S. L., Piñeiro, R., Heuertz, M., Duminil, J., Daïnou, K., Doucet, J.-L., et al. (2020). Population genomics of the widespread African savannah trees Afzelia africana and Afzelia quanzensis reveals no significant past fragmentation of their distribution ranges. *Am. J. Bot.* 107, 498–509. doi: 10.1002/ajb2.1449

Donoghue, M. J., and Edwards, E. J. (2014). Biome shifts and niche evolution in plants. *Ann. Rev. Ecol. Evol. Syst.* 45, 547–572. doi: 10.1146/annurev-ecolsys-120213-091905

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

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian P hylogenetics with BEAUti and the BEAST 1.7 Research article. *Soc. Mol. Biol. Evol.* 29, 1969–1973. doi: 10.1093/molbev/mss075

Duminil, J., Kenfack, D., Viscosi, V., Grumiau, L., and Hardy, O. J. (2012). Testing species delimitation in sympatric species complexes: the case of an African tropical tree, *Carapa* spp.(Meliaceae). *Mol. Phylogenet. Evol.* 62, 275–285. doi: 10.1016/j.ympev.2011.09.020

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

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379

Escudero, M., Eaton, D. A. R., Hahn, M., and Hipp, A. L. (2014). Genotyp- ing-By-Sequencing as a tool to infer phylogeny and ancestral hybridization: a case study in Carex (Cyperaceae). *Mol. Phylogenet. Evol.* 79, 359–367. doi: 10.1016/j.ympev.2014.06.026

Esselstyn, J. A., Evans, B. J., Sedlock, J. L., Anwarali Khan, F. A., and Heaney, L. R. (2012). Single-locus species delimitation: a test of the mixed Yule–coalescent model, with an empirical application to Philippine round-leaf bats. *Proc. R. Soc. B Biol. Sci.* 279, 3678–3686. doi: 10.1098/rspb.2012.0705

Ezard, T., Fujisawa, T., and Barraclough, T. (2013). *R package splits: SPecies' LImits by Threshold Statistics, version 1.0-18/r45*. Available online at: http://r-forge.r-project.org/projects/splits/ (accessed September, 2019).

Fernández-Mazuecos, M., Mellers, G., Vigalondo, B., Sáez, L., Vargas, P., and Glover, B. J. (2018). Resolving recent plant radiations: power and robustness of genotyping-by-sequencing. *Syst. Biol.* 67, 250–268. doi: 10.1093/sysbio/syx062

Freitas, C. G., Bacon, C. D., Souza-Neto, A. C., and Collevatti, R. G. (2019). Adjacency and area explain species bioregional shifts in neotropical palms. *Front. Plant Sci.* 10:55. doi: 10.3389/fpls.2019.00055

Fujisawa, T., and Barraclough, T. G. (2013). Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: a revised method and evaluation on simulated data sets. *Syst. Biol.* 62, 707–724. doi: 10.1093/sysbio/syt033

Heled, J., and Drummond, A. J. (2015). Calibrated birth–death phylogenetic time-tree priors for bayesian inference. *Syst. Biol.* 64, 369–383. doi: 10.1093/sysbio/syu089

Heuertz, M., Caron, H., Scotti-Saintagne, C., Pétronelli, P., Engel, J., Tysklind, N., et al. (2020). The hyperdominant tropical tree Eschweilera coriacea

(Lecythidaceae) shows higher genetic heterogeneity than sympatric Eschweilera species in French Guiana. *Plant Ecol. Evol.* 153, 67–81. doi: 10.5091/plecevo.2020.1565

Heuertz, M., Duminil, J., Dauby, G., Savolainen, V., and Hardy, O. J. (2014). Comparative phylogeography in rainforest trees from Lower Guinea. Africa. *PLoS One* 9:e84307. doi: 10.1371/journal.pone.0084307

Hipp, A. L., Eaton, D. A. R., Cavender-Bares, J., Fitzek, E., Nipper, R., and Manos, P. S. (2014). A framework phylogeny of the American oak clade based on sequenced RAD data. *PLoS One* 9:e93975. doi: 10.1371/journal.pone.0093975

Holstein, N., and Renner, S. S. (2011). A dated phylogeny and collection records reveal repeated biome shifts in the African genus Coccinia (Cucurbitaceae). *BMC Evol. Biol.* 11:28. doi: 10.1186/1471-2148-11-28

Ikabanga, D. U., Stévart, T., Koffi, G. K., Monthe, F. K., Doubindou, N. E. C., Dauby, G., et al. (2017). Combining morphology and population genetic analysis uncover species delimitation in the widespread African tree genus Santiria (Burseraceae). *Phytotaxa* 321, 166–180.

Institut National pour l'Etude Agronomique du Congo-belge [INEAC] (1952). *Spermatophytes. Flore du Congo Belge et du Ruanda-Urundi*. Bruxelles: INEAC, 579.

International Union for Conservation of Nature and Natural Resources [IUCN] (2012). *IUCN Red List of Threatened Species*. Available online at: www.iucnredlist.org (accessed January 08, 2012).

Jacobs, B. F. (2004). Palaeobotanical studies from tropical Africa: relevance to the evolution of forest, woodland and savannah biomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 359, 1573–1583. doi: 10.1098/rstb.2004.1533

Jacobs, B. F., Pan, A. D., and Scotese, C. R. (2010). *A review of the Cenozoic vegetation history of Africa. Cenozoic Mammals of Africa*. Berkeley: University of California Press, 57–72.

Kadiri, A. B., and Olowokudejo, J. D. (2008). Comparative foliar epidermal morphology of the West African species of the genus Afzelia Smith (Leguminosae: Caesalpinioideae). *Gayana Bot.* 65, 84–92.

Koffi, K. G., Heuertz, M., Doumenge, C., Onana, J. M., Gavory, F., and Hardy, O. J. (2010). A combined analysis of morphological traits, chloroplast and nuclear DNA sequences within Santiria trimera (Burseraceae) suggests several species following the Biological Species Concept. *Plant Ecol. Evol.* 143, 160–169. doi: 10.5091/plecevo.2010.433

Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. N. -M., and Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Syst. Biol.* 64, 1032–1047. doi: 10.1093/sysbio/syv053

Léonard, J. J. G. (1950). Notes sur les genres paleotropicaux Afzelia, Intsia et Pahudia (Legum. Caesalp.). *Reinwardtia* 1, 61–66.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50, 913–925. doi: 10.1080/106351501753462876

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lischer, H. E. L., and Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28, 298–299. doi: 10.1093/bioinformatics/btr642

Lissambou, B-J., Hardy, O. J., Atteke, C., Stevart, T., Dauby, G., Mbatchi, B., et al. (2018). Taxonomic revision of the African genus Greenwayodendron (Annonaceae). *Phytokeys* 114, 55–93. doi: 10.3897/phytokeys.114.27395

LPWG (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon* 66, 44–77.

Maurin, O., Davies, T. J., Burrows, J. E., Daru, B. H., Yessoufou, K., Muasya, A.M., et al., (2014). Savanna fire and the origins of the 'underground forests' of Africa. *New Phytol.* 204, 201–214. doi: 10.1111/nph.12936

Miller, M. A. (2009). *The CIPRES Portals. CIPRES. Arch. By WebCite(r)*. Available online at: https://www.webcitation.org/5imQlJeQa (accessed December, 2019).

Miller, C. S., and Gosling, W. D. (2014). Quaternary forest associations in lowland tropical West Africa. *Quat. Sci. Rev.* 84, 7–25. doi: 10.1016/j.quascirev.2013.10.027

Monthe, F. K., Migliore, J., Duminil, J., Bouka, G., Demenou, B. B., Doumenge, C., et al. (2019). Phylogenetic relationships in two African Cedreloideae tree genera (Meliaceae) reveal multiple rain/dry forest transitions. *Perspect. Plant Ecol. Evol. Syst.* 37, 1–10. doi: 10.1016/j.ppees.2019.01.002

Morley, R. J. (2000). *Origin and Evolution of Tropical Rain Forests*. New York, NY: John Wiley & Sons.

Nicotra, A. B., Chong, C., Bragg, J. G., Ong, C. R., Aitken, N. C., Chuah, A., et al. (2016). Population and phylogenomic decomposition via genotyping-by-sequencing in Australian Pelargonium. *Mol. Ecol.* 25, 2000–2014. doi: 10.1111/mec.13584

Pan, A. D., Jacobs, B. F., and Herendeen, P. S. (2010). Detarieae *sensu lato* (Fabaceae) from the Late Oligocene (27.23 Ma) Guang River flora of northwestern Ethiopia. *Bot. J. Linn. Soc.* 163, 44–54. doi: 10.1111/j.1095-8339.2010.01044.x

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Pennington, R. T., and Hughes, C. E. (2014). The remarkable congruence of New and Old World savannah origins. *New Phytol.* 204, 4–6. doi: 10.1111/nph.12996

Pennington, R. T., and Lavin, M. (2016). The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *New Phytol.* 210, 25–37. doi: 10.1111/nph.13724

Pickrell J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* e1002967. doi: 10.1371/journal.pgen.1002967

Pinheiro, F., De Barros, F., Palma-Silva, C., Meyer, D., Fay, M. F., Suzuki, R. M., et al. (2010). Hybridization and introgression across different ploidy levels in the Neotropical orchids Epidendrum fulgens and *E. puniceoluteum* (Orchidaceae). *Mol. Ecol.* 19, 3981–3994. doi: 10.1111/j.1365-294x.2010.04780.x

Plana, V. (2004). Mechanisms and tempo of evolution in the African Guineo-Congolian rainforest. *Philos. Trans. R. Soc. B Biol. Sci.* 359, 1585–1594. doi: 10.1098/rstb.2004.1535

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., et al. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol.* 55, 595–609. doi: 10.1080/10635150600852011

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

Rambaut, A. (2007). *FigTree, a Graphical Viewer of Phylogenetic Trees*. Available online at: http://tree.bio.ed.ac.uk/software/Figtree/ (accessed March, 2020).

Rambaut, A., and Drummond, A. J. (2016). *Tracer v 1.6*. Scotland: University of Edinburg, 2007.

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. . doi: 10.7717/peerj.2584

Sack, L., and Scoffoni, C. (2013). Leaf venation: structure, function, development, evolution, ecology and applications in the past, present and future. *New Phytol.* 198, 983–1000. doi: 10.1111/nph.12253

Salzmann, U., and Hoelzmann, P. (2005). The Dahomey Gap: an abrupt climatically induced rain forest fragmentation in West Africa during the late Holocene. *Holocene* 15, 190–199. doi: 10.1191/0959683605hl799rp

Sarnthein, M., and Fenner, J. (1988). Global wind-induced change of deep-sea sediment budgets, new ocean production and CO2 reservoirs ca. 3.3–2.35 Ma BP. *Philos. Trans. R. Soc. London, Ser. A* 318, 487–504. doi: 10.1098/rstb.1988.0020

Satabié, B. (1994). Biosystématique et Vicariance dans la flore Camerounaise. *Bull. Jard. Bot. Nat. Belg.* 63, 125–170.

Senut, B., Pickford, M., and Ségalen, L. (2009). Neogene desertification of Africa. *Comptes Rendus. Geosci.* 341, 591–602. doi: 10.1016/j.crte.2009.03.008

Simon, M. F., Grether, R., de Queiroz, L. P., Skema, C., Pennington, R. T., and Hughes, C. E. (2009). Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20359–20364. doi: 10.1073/pnas.0903410106

Simon, M. F., and Pennington, R. T. (2012). The evolution of adaptations of woody plants in the savannas of the Brazilian cerrado. *Int. J. Plant Sci.* 173, 711–723.

Soltis, D. E., Soltis, P. S., Schemske, D. W., Hancock, J. F., Thompson, J. N., Husband, B. C., et al. (2007). Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* 56, 13–30.

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

Stange, M., Sanchez-Villagra, M. R., Salzburger, W., and Matschiner, M. (2018). Bayesian divergence-time estimation with genome-wide single-nucleotide polymorphism data of sea catfishes (Ariidae) supports Miocene closure of the Panamanian Isthmus. *Syst. Biol.* 6, 681–699. doi: 10.1093/sysbio/syy006

Tng, D. Y. P., Jordan, G. J., and Bowman, D. M. J. S. (2013). Plant traits demonstrate that temperate and tropical giant eucalypt forests are ecologically convergent with rainforest not savanna. *PLoS One* 8:e0084378. doi: 10.1371/journal.pone.0084378

Tosso, F., Hardy, O. J., Doucet, J. L., Daïnou, K., Kaymak, E., and Migliore, J. (2018). Evolution in the Amphi- Atlantic tropical genus Guibourtia (Fabaceae, Detarioideae), combining NGS phylogeny and morphology. *Mol. Phylogenet. Evol.* 120, 83–93. doi: 10.1016/j.ympev.2017.11.026

Veranso-Libalah, M. C., Kadereit, G., Stone, R. D., and Couvreur, T. L. P. (2018). Multiple shifts to open habitats in Melastomateae (Melastomataceae) congruent with the increase of African Neogene climatic aridity. *J. Biogeogr.* 45, 1420–1431. doi: 10.1111/jbi.13210

Wiens, J. J., and Donoghue, M. J. (2004). Historical biogeography, ecology, and species richness. *Trends Ecol. Evol.* 19, 639–644. doi: 10.1016/j.tree.2004.09.011

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2007). The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13875–13879. doi: 10.1073/pnas.0811575106

Zachos, J. C., Dickens, G. R., and Zeebe, R. E. (2008). An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. *Nature* 451, 279–283. doi: 10.1038/nature06588

Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 29, 2869–2876. doi: 10.1093/bioinformatics/btt499

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. F.R.S. *Philos. Trans. R. Soc. B Biol. Sci.* 213, 21–87. doi: 10.1098/rstb.1925.0002

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: 26 June 2020
doi: 10.3389/fpls.2020.00819

Museomics Unveil the Phylogeny and Biogeography of the Neglected Juan Fernandez Archipelago *Megalachne* and *Podophorus* Endemic Grasses and Their Connection With Relict Pampean-Ventanian Fescues

*María Fernanda Moreno-Aguilar[1], Itziar Arnelas[2], Aminael Sánchez-Rodríguez[2], Juan Viruel[3] and Pilar Catalán[1,4,5]**

[1] Escuela Politécnica Superior de Huesca, Universidad de Zaragoza, Huesca, Spain, [2] Departamento de Ciencias Biológicas, Universidad Técnica Particular de Loja, Loja, Ecuador, [3] Royal Botanic Gardens, Kew, Richmond, United Kingdom, [4] Grupo de Bioquímica, Biofísica y Biología Computacional (BIFI, UNIZAR), Unidad Asociada al CSIC, Zaragoza, Spain, [5] Department of Botany, Institute of Biology, Tomsk State University, Tomsk, Russia

OPEN ACCESS

Edited by:
Nina Rønsted,
National Tropical Botanical Garden,
United States

Reviewed by:
Martin Röser,
Martin Luther University
of Halle-Wittenberg, Germany
Josef Greimler,
University of Vienna, Austria
Guillaume Besnard,
UMR 5174 Evolution et Diversite
Biologique (EDB), France

***Correspondence:**
Pilar Catalán
pcatalan@unizar.es

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

Received: 31 January 2020
Accepted: 22 May 2020
Published: 26 June 2020

Citation:
Moreno-Aguilar MF, Arnelas I,
Sánchez-Rodríguez A, Viruel J and
Catalán P (2020) Museomics Unveil
the Phylogeny and Biogeography
of the Neglected Juan Fernandez
Archipelago Megalachne
and Podophorus Endemic Grasses
and Their Connection With Relict
Pampean-Ventanian Fescues.
Front. Plant Sci. 11:819.
doi: 10.3389/fpls.2020.00819

Oceanic islands constitute natural laboratories to study plant speciation and biogeographic patterns of island endemics. Juan Fernandez is a southern Pacific archipelago consisting of three small oceanic islands located 600–700 km west of the Chilean coastline. Exposed to current cold seasonal oceanic climate, these 5.8–1 Ma old islands harbor a remarkable endemic flora. All known Fernandezian endemic grass species belong to two genera, *Megalachne* and *Podophorus*, of uncertain taxonomic adscription. Classical and modern classifications have placed them either in Bromeae (*Bromus*), Duthieinae, Aveneae/Poeae, or Loliinae (fine-leaved *Festuca*); however, none of them have clarified their evolutionary relationships with respect to their closest *Festuca* relatives. *Megalachne* includes four species, which are endemic to Masatierra (Robinson Crusoe island) (*M. berteroniana* and *M. robinsoniana*) and to Masafuera (Alejandro Selkirk island) (*M. masafuerana* and *M. dantonii*). The monotypic *Podophorus bromoides* is a rare endemic species to Masatierra which is only known from its type locality and is currently considered extinct. We have used museomic approaches to uncover the challenging evolutionary history of these endemic grasses and to infer the divergence and dispersal patterns from their ancestors. Genome skimming data were produced from herbarium samples of *M. berteroniana* and *M. masafuerana*, and the 164 years old type specimen of *P. bromoides*, as well as for a collection of 33 species representing the main broad- and fine-leaved Loliinae lineages. Paired-end reads were successfully mapped to plastomes and nuclear ribosomal cistrons of reference *Festuca* species and used to reconstruct phylogenetic trees. Filtered ITS and *trn*TLF sequences from these genomes were further combined with our large Loliinae data sets for accurate biogeographic reconstruction. Nuclear and plastome data recovered a strongly supported fine-leaved Fernandezian clade where *Podophorus* was

resolved as sister to *Megalachne*. Bayesian divergence dating and dispersal-extinction-cladogenesis range evolution analyses estimated the split of the Fernandezian clade from its ancestral southern American Pampas-Ventanian Loliinae lineage in the Miocene-Pliocene transition, following a long distance dispersal from the continent to the uplifted volcanic palaeo-island of Santa Clara-Masatierra. Consecutive Pliocene-Pleistocene splits and a Masatierra-to-Masafuera dispersal paved the way for *in situ* speciation of *Podophorus* and *Megalachne* taxa.

Keywords: ancestral range reconstruction, endemic Loliinae grasses, Fernandezian clade, genome skimming, phylogenomics, taxonomically neglected species

INTRODUCTION

Genomic data are increasingly called upon to elucidate evolutionary and taxonomic challenges posed by several cryptic or ambiguously related organisms, which could not be resolved using traditional approaches, such as morphometrics or standard molecular methods (Harrison and Kidner, 2011; Straub et al., 2012; Carter et al., 2020; Larridon et al., 2020). The advent of the high-throughput sequencing (HTS) methods have outpaced classical molecular barcoding and phylogenetic procedures based on few molecular markers that have served to build phylogenies with constrained resolution limits (Diaz-Perez et al., 2018; Sancho et al., 2018). While the results obtained from the genomic-based approaches are overall congruent with previous findings based on reduced sets of genes and genetic markers (Saarela et al., 2018), the thoroughly dissection of genomes have untapped large sets of taxonomically informative gene copy variants or single nucleotide polymorphism (SNPs) and have allowed the reconstruction of better resolved and more strongly supported phylogenies (Soltis et al., 2018). These new metadata have facilitated the identification of previously neglected cryptic taxa (Spriggs et al., 2019) and the construction of more robust phylogenetic trees where the evolutionary positions of previously unknown, doubtful, or ambiguous lineages have been elucidated in some cases (Leebens-Mack et al., 2019; Li et al., 2019).

The application of HTS methods to the analysis of museum collections, defined as museomics, has revolutionized the study of the organismic diversity (Besnard et al., 2014; Nevill et al., 2020). Plant herbarium specimens were occasionally used in traditional phylogenetic and population genetic studies due to the poor preservation of the specimens or their low quality DNA. Herbarium specimens have been progressively incorporated to taxonomic and evolutionary studies using HTS methods thanks to the simultaneous generation of a large quantity of sequences for the different genomes present in an organism (Straub et al., 2012; Besnard et al., 2014). Among the HTS approaches used with both herbarium and fresh collections, genome skimming (Dodsworth, 2015; Richter et al., 2015) has been successfully applied to reconstruct DNA genomes and regions that exist in multiple copies, such as plastomes, mitomes and the nuclear ribosomal cistron, and even some nuclear single copy genes (Besnard et al., 2014). Among other advances, museomics has untapped the placement of recently extinct taxa in phylogenies (Sebastian et al., 2010; Welch et al., 2016; Zedane et al., 2016; Silva et al., 2017). Thus, the combined use of current and extinct plant species samples, and of herbarium and recently collected samples allows to uncover largely sampled phylogenetic trees of plant lineages (Malakasi et al., 2019).

Oceanic archipelagos have been recognized as hotspots of diversity and natural laboratories for long distance colonization and plant speciation events (Triantis et al., 2016). Juan Fernandez is one of the smallest oceanic archipelagos. It consists of three small islands located in the southern Pacific, 580–730 km offshore of the western Chilean coast [Masatierra or Robinson Crusoe (47.94 km^2, 0–915 masl), Masafuera or Alejandro Selkirk (49.52 km^2, 0–1,319 masl), and Santa Clara (2.21 km^2, 0–350 masl)] (Stuessy et al., 1992, 2017). The two main islands have similar sizes but differ in plant communities and diverse grassland extensions due to their different ages and erosional patterns (Greimler et al., 2017), and are separated each other by 181 km (**Figure 1**). The current Fernandezian volcanic islands are relatively young (Stuessy et al., 1984). Despite its total small area (100.2 km^2), the archipelago harbors one of the richest endemic floras (60% vascular species, 11% genera, 1 paleoherb family; Stuessy et al., 1992). Floristic studies indicate that 55 grass species grow in Juan Fernandez archipelago; most of them are invasive taxa except five endemic species that belong to the Fernandezian *Megalachne* Steud. and *Podophorus* Phil. genera (Baeza et al., 2007; Peña et al., 2017; Penneckamp-Furniel and Villegas, 2019). *Megalachne* and the monotypic genus *Podophorus* have been historically assigned to different temperate grass tribes. *Megalachne* was originally described by Steudel in 1854 as close to *Bromus* (it was also described as *Pathantera* by Philippi in 1856), though they differ in the number and disposition of the stigmas (three apical in *Megalachne*, two subapical in *Bromus*) and the shape of the glume apex (aristulate in *Megalachne*, mutique in *Bromus;* Hackel, 1887). However, Pilger in 1920 and Skottsberg in 1922 transferred, respectively, *Megalachne* and *Pathantera* to *Bromus*, based on the sharing of laterally compressed spikelets and keeled lemmas, such as those in *Bromus* sect. *Ceratochloa* (Peña et al., 2017) thus classifying them within tribe Bromeae. In 1954, Pilger recognized *Megalachne* as a separate genus from *Bromus* (Peña et al., 2017); Tateoka (1962) using evidences from the morphology and apical hairiness of the ovary, apical emergence of stigmas, and type of starch grains and serology, suggested the proximity of *Megalachne* to *Festuca*, thus attributing it to tribe Poeae (subtribe Loliinae). The taxonomic adscription of *Megalachne* and *Podophorus* to

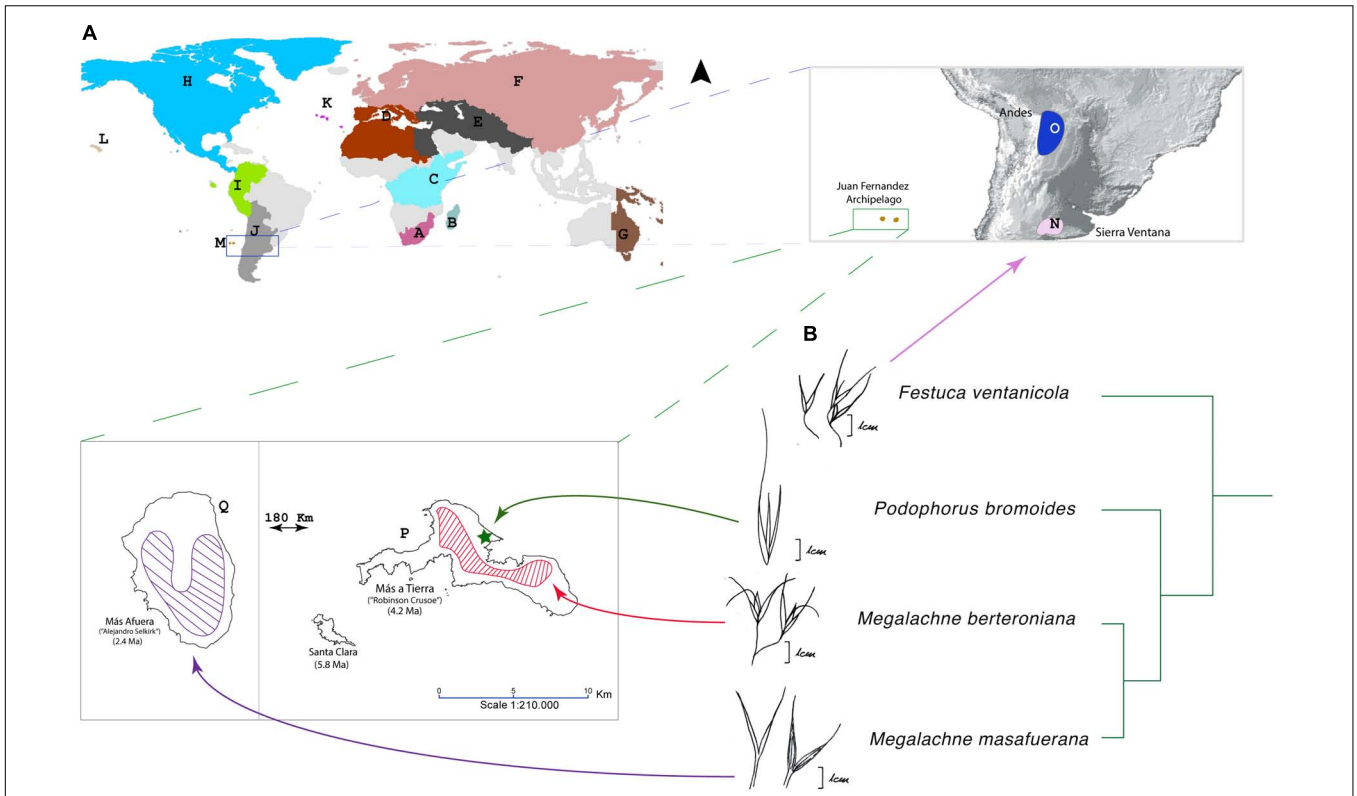

FIGURE 1 | (A) Distribution map of the Juan Fernandez archipelago *Megalachne* and *Podophorus* species included in this study [*M. berteroniana* and *M. masafuerana* distributions have been obtained from Peña et al. (2017) and Penneckamp-Furniel and Villegas (2019); *P. bromoides* is tentatively mapped in the Cumberley bay of Masatierra; cf. Baeza et al. (2007)] and of the Pampean-Ventanian *Festuca ventanicola*, showing the world and the American-Vulpia-Pampas Operational areas (OAs) used in the biogeographical analyses of Loliinae. OAs: Loliinae DEC model: A – South Africa; B – Madagascar and Mascarenes; C – Tropical Africa; D – Mediterranean; E – Irano-Turanian-Himalayan region; F – Eurasia (Eurosiberian region); G – Australasia (New Zealand, south-western Australia, Papua-New Guinea); H – North and Central America; I – Northern Andes; J – Southern Andes and Southern South America; K – Macaronesia; L – Hawaii; M – Juan Fernández; American-Vulpia-Pampas DEC model: H – North and Central America; N – Pampas-Ventania; O – Andes; P – Masatierra; Q – Masafuera. **(B)** Phylogenetic ITS-TLF subtree showing the relationships of the Fernandezian *Podophorus bromoides*, *Megalachne berteroniana*, and *M. masafuerana* grasses and its close relative *Festuca ventanicola*; drawns of the floral "vulpioid" phenotype are shown for each species.

tribe Poeae was accepted in most grass classifications though Soreng et al. (2003) assigned them initially to tribe Stipeae subtribe Duthieinae based on the overall habit resemblance. Nonetheless, the comprehensive morphological and molecular study of the newly delimited tribe Duthieeae of Schneider et al. (2011) demonstrated, using ITS sequences, that *Megalachne* and *Podophorus* were not part of this early diverging pooid lineage, and suggested that they likely belonged to the Aveneae/Poeae complex. In recent studies, phylogenetic analyses conducted by Schneider et al. (2012) and Tkach et al. (2020) using, respectively, nuclear ITS and plastid *mat*K sequences and nuclear ITS and ETS and plastid *mat*K, *trn*K, and *trn*LF sequences corroborated it, showing that *Megalachne* was nested within the fine-leaved Loliinae clade.

Megalachne and *Podophorus* differentiate from each other in the number of florets per spikelet [3–6 in *Megalachne*, 1–(+1 sterile) in *Podophorus*), the type of lemma (keeled vs. rounded], the length of the glumes (equal vs. shorter than anthecium) and the prolongation of the rachilla apex (shorther vs. equal than anthecium; Baeza et al., 2007; Kellogg, 2015; Peña et al., 2017). *Megalachne* consisted until recently of two species, *M. berteroniana* Steud. and *M. masafuerana* (Skottsb. & Pilg. ex. Pilg.) Matthei, endemic to the Masatierra and Masafuera islands, respectively (Baeza et al., 2007). Both species grow in coastal and mountain cliffs in their respective islands (Danton et al., 2006). Recent morphological studies have identified two new species, *M. robinsoniana* C. Peña, endemic to Masatierra (Peña et al., 2017), and *M. dantonii* Penneck. & Gl. Rojas, endemic to Masafuera (Penneckamp-Furniel and Villegas, 2019). The four *Megalachne* species differ in the lengths of the lemma and glume awns and the number of florets per spikelet (Peña et al., 2017; Penneckamp-Furniel and Villegas, 2019). The systematic and evolutionary fate of *Podophorus* is more enigmatic. Its single species *P. bromoides* Phil. is only known from its type specimens, collected in Masatierra and described by Philippi in 1856, and is currently considered to be extinct (Baeza et al., 2007).

Loliinae is one of the largest subtribes of the temperate pooid grasses and contains pasture and forage species of high ecological and economic importance. Its largest genus *Festuca*

is formed by ∼600 worldwide distributed species inhabiting cool seasonal regions of both hemispheres and high tropical mountains (Catalan, 2006). Molecular phylogenetic studies have shown that *Festuca* is largely paraphyletic (Catalan et al., 2007; Inda et al., 2008; Minaya et al., 2017). Recent studies, based on the Schneider et al. (2011, 2012) ITS and *mat*K data and previous morphological findings, reclassified *Megalachne* and *Podophorus* within subtribe Loliinae (Soreng et al., 2015, 2017), however, they did not identify the closest relatives of these Fernandezian grasses. The phylogenetic relationships obtained by previous authors for the three studied *Megalachne* and *Podophorus* taxa were also taxonomically incongruent, showing a closer relationship of *M. berteroniana* to *P. bromoides* than to its congener *M. masafuerana* (Schneider et al., 2011).

Here we have used a museomic approach based on genome skimming data to uncover the phylogenetic and biogeographical history of the neglected Fernandezian *Megalachne* and *Podophorus* grasses. The aims of our study were to (i) infer the phylogeny of *Megalachne* and *Podophorus* within a large sample representation of Loliinae lineages; (ii) identify the closest relatives of the Fernandezian grasses; (iii) reconstruct the relationships among the *Megalachne* and *Podophorus* taxa; (iv) estimate divergence times of the Fernandezian lineages; and (v) infer the colonization patterns and speciation events of the ancestors of *Megalachne* and *Podophorus* in the Juan Fernandez islands.

MATERIALS AND METHODS

Sampling

Representative samples of *Megalachne*, *Podophorus*, and other Loliinae genera were included in the study (**Figure 1** and **Table 1**). Herbarium samples of *Megalachne berteroniana* and *M. masafuerana* provided by the Oregon State University Herbarium (OSC11751 and OSC9150 collections; **Table 1**) were used to isolate high quality and quantity DNA for genome sequencing and downstream evolutionary analyses. A herbarium sample of *M. robinsoniana* provided by the Concepción University Herbarium (CONC40598 collection) failed to generate good quality DNA for the study. The recently described *M. dantonii* species (Penneckamp-Furniel and Villegas, 2019) could not be included in our study. A 164 years old sample of the currently considered extinct *Podophorus bromoides* Phil., only known from its three type specimens, was provided by the Royal Botanic Gardens Kew's Herbarium (Philippi 1861, isotype collection; **Table 1**)[1] and was successfully used for genome skimming sequencing and downstream analysis. In our aim to identify the closest relatives of the Fernandezian *Megalachne* and *Podophorus* grasses, DNA was also isolated from 33 Loliinae samples (**Table 1**) representing all the known broad-leaved, intermediate, and fine-leaved Loliinae lineages (Inda et al., 2008; Minaya et al., 2017) and used for genome skimming sequencing and phylogenomic analyses. Some of these samples were collected from poorly explored geographical areas, including four new *Festuca* samples from South America, the putative region of origin of the ancestors of *Megalachne* and *Podophorus* (Stuessy et al., 2017) and two from Tropical Africa and South Africa. In addition, two new Loliinae samples from South America and one sample from South Africa not studied before (**Table 1**) were sequenced for the nuclear ITS (ITS1-5.8S-ITS2) and the plastid *trn*TL (*trn*T-*trn*L intergenic spacer) and *trn*LF (*trn*A-Leu, *trn*L-*trn*F intergenic spacer, *trn*A-Phe) loci, together with 97 samples from a wide-sampling of all currently known Loliinae lineages (Inda et al., 2008; Minaya et al., 2017). Although species of *Megalachne* and *Podophorus* and other fine-leaved Loliinae genera have been synonymized to *Festuca*, and those of broad-leaved Loliinae to different festucoid genera in recent studies (Soreng et al., 2017; Tkach et al., 2020), we follow the *Festuca* sensu lato classification of Catalan et al. (2007) which is based on an evolutionary systematic criterion that is nomenclaturally conservative and maintains a paraphyletic *Festuca* (with subgenera and sections) and other traditionally recognized genera until more complete phylogenetic studies of Loliinae are conducted. We have selected this scenario because of present uncertainties about the phylogeny of several Loliinae lineages and taxonomic and nomenclatural instability of the *Festuca* sensu stricto (i.e., fine-leaved Loliinae lineages) classification, that would leave some broad-leaved Loliinae lineages without name or with unclear adscription (e.g., some broad-leaved "*Festuca*"). It could be possible, however, that genera phylogenetically embedded within the large Loliinae clade or its fine-leaved subclade would be subordinated to *Festuca*, once all or most of the Loliinae taxa are phylogenetically analyzed and consistent synapomorphies are defined. At this respect, nomenclatural combinations have been proposed for the fine-leaved *Megalachne* (*Festuca megalachna* Röser & Tkach; *F. masafuerana* (Skottsb. & Pilg. ex Pilg.) Röser & Tkach; *F. robinsoniana* (C.M.Peña) Röser & Tkach; *F. dolichathera* Röser & Tkach) and *Podophorus* (*F. masatierrae* Röser & Tkach) species synonymized to *Festuca* (Tkach et al., 2020). Sixteen additional species were added as outgroups to provide reliable fossil calibration points for molecular dating (**Supplementary Table S1**).

DNA Extraction and Sequencing

The 36 samples used in this study were obtained from herbarium specimens (AARHUS, K, MO, US, OS, CONC, HUTPL, University of Zaragoza), silica gel dried leaf tissues collected in field trips, and fresh leaves collected from plants growing in the Universidad de Zaragoza – Escuela Politécnica Superior de Huesca common garden (**Table 1** and **Supplementary Table S1**). Total DNA from fresh and silica gel dried samples was isolated following the DNeasy Plant Mini kit (Qiagen, Valencia, CA, United States) protocol using 20–30 mg of dry leaf tissue or 20 mg of fresh tissue ground to powder with liquid nitrogen. Total DNA from herbarium samples was extracted using a modified CTAB protocol (Doyle and Doyle, 1987) using ∼20 mg of tissue. DNA concentration was quantified with a Qubit fluorometer (Invitrogen by Life Technologies, Carlsbad, California, United States) and DNA quality was evaluated with Biodrop (Harvard Bioscience). The integrity of the DNA was

[1]https://apps.kew.org/herbcat/getImage.do?imageBarcode=K000433684

TABLE 1 | List of taxa included in the phylogenomic study of the Fernandezian and other Loliinae grasses.

Taxon	Source	Ploidy	No. reads	Genbank/Phytozome accession No.	
				Plastome	rDNA cistron
Festuca abyssinica	Tanzania: Kilimanjaro	4x	12041	SAMN14647043	MT145276
Festuca africana	Uganda: Bwindi forest	10x	13549	SAMN14647044	MT145277
Festuca amplissima	Mexico: Barranca del Cobre	6x	12058	SAMN14647045	MT145278
Festuca arundinacea var. *letourneuxiana*	Morocco: Atlas Mountains	10x	16839	SAMN14647059	MT145292
Festuca asplundii	Ecuador: Saraguro	6x	25088	SAMN14647046	MT145279
Festuca caldasii	Ecuador: Las Chinchas -Tambara	?	9863	SAMN14647047	MT145280
Festuca capillifolia	Spain: Cazorla	2x	13430	SAMN14647048	MT145281
Festuca chimborazensis	Ecuador: Chimborazo-Cotopaxi	4x	10913	SAMN14647049	MT145282
Festuca durandoi	Portugal: Alto do Espinheiro	2x	12688	SAMN14647050	MT145283
Festuca eskia	Spain: Picos de Europa	2x	24041	SAMN14647051	MT145284
Festuca fenas	Spain: Madrid	4x	16112	SAMN14647052	MT145285
Festuca fimbriata	Argentina: Apóstoles	6x	15741	SAMN14647053	MT145286
Festuca fontqueriana	Morocco: Rif, Outa-El-Kadir	2x	22187	SAMN14647054	MT145287
Festuca gracillima	Argentina: Tierra de Fuego	6x	13888	SAMN14647055	MT145288
Festuca holubii	Ecuador: Saraguro	?	10264	SAMN14647056	MT145289
Festuca francoi	Portugal: Azores	2x	17592	SAMN14647057	MT145290
Festuca lasto	Spain: Los Alcornocales	2x	21581	SAMN14647058	MT145291
Festuca mairei	Morocco: Atlas Mountains	4x	19134	SAMN14647060	MT145293
Festuca molokaiensis	United States: Hawaii, Molokai	?	12188	SAMN14647061	MT145294
Festuca ovina	Russia: Gatchinskii Raion	2x	11364	SAMN14647062	MT145295
Festuca pampeana	Argentina: Sierra Ventana	6x	14862	SAMN14647063	MT145296
Festuca paniculata	Spain: Puerto de los Castaños	2x	35808	SAMN14647064	MT145297
Festuca parvigluma	China: Baotianman	4x	15872	SAMN14647065	MT145298
Festuca pratensis	England: USDA/283306	2x	30021	SAMN14647066	MT145301
Festuca procera	Ecuador: Riobamba	4x	12189	SAMN14647067	MT145299
Festuca pyrenaica	Spain: Pyrenees, Tobacor	4x	40669	SAMN14647068	MT145300
Festuca pyrogea	Argentina: Tierra de fuego	?	16835	SAMN14647069	MT145302
Festuca quadridentata	Ecuador: Chimborazo	?	15091	SAMN14647070	MT145303
Festuca spectabilis	Bosnia-Hercegovina: Troglav	6x	12960	SAMN14647071	MT145304
Festuca superba	Argentina: Jujuy, Yala	8x	12193	SAMN14647072	MT145305
Festuca triflora	Morocco: Rif, Ketama	2x	24472	SAMN14647073	MT145306
Megalachne berteroniana	Chile: JuanFernandez, Masatierra	?	5288	SAMN14647074	MT145307
Megalachne masafuerana	Chile: JuanFernandez, Masafuera	?	6134	SAMN14647075	MT145308
Podophorus bromoides	Chile: JuanFernandez, Masatierra	?	6694	SAMN14668162	–
Vulpia ciliata	Spain: Mar de Ontígola	4x	11801	SAMN14647076	MT145309
Vulpia sicula	Italia: Sicilia, Madone	2x	11327	SAMN14647077	MT145310
Outgroups					
Brachypodium distachyon	Iraq: near Salakudin	2x	–	NC_011032.1	phytozome.jgi.doe.gov, Bd21 v.3.1
Oryza sativa subsp. *japonica*	cv. PA64S; cv. Nipponbare	2x	–	AY522331.1	AP008215

Taxon, source, ploidy level, number of Illumina reads, and plastome and rDNA cistron Genbank codes are indicated for each sample. Newly generated Loliinae plastome data and sequences have been deposited in Genbank under BioProject PRJNA626668 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA626668) and at Github [https://github.com/Bioflora/Loliinae_plastomes (unfiltered, filtered) and https://github.com/Bioflora/Podophorus_plastome (unfiltered, filtered)].

further checked in a 1% agarose gel. Overall the qualities and quantities of the DNAs were appropriate for genome skimming (∼5 μg, 50 ng/μl), except that of *P. bromoides*, which had <1 ng/μl.

DNAs obtained from three *Megalachne* and *Podophorus* samples plus 33 Loliinae samples were used to construct a genomic library for shotgun sequencing using Illumina technology. The library from freshly and herbarium collected materials DNAs was prepared with KAPA Hyper Prep Kit for PCR-free workflows (Roche Kapa Biociences) with some minor modifications. In brief, 1.0 μg of genomic DNA was sheared in a Covaris^{TM} E220 focused-ultrasonicator into Covaris microTUBE AFA Fiber Pre-Slit Snap-Cap tubes with the following parameters: sample volume 55 μl, duty cycle 15%, intensity 450, cycles/burst 200, time 100 s, temperature 4°C, in order to reach the fragment sizes of ∼200–400 bp.

The sheared DNA was end-repaired, adenylated and ligated to IDT adaptors with unique dual-matched indexes (Integrated DNA Technologies) for paired end sequencing. The adaptor-modified end library was size selected and purified with AMPure XP beads (Agencourt, Beckman Coulter) in order to eliminate non-ligated adapters and adapter dimers. Final library size was confirmed on an Agilent 2100 Bioanalyzer with the DNA 7500 assay. The *Podophorus bromoides* library yielded 13 ng/μl and two normally distributed fragment size distributions of 200 and 500 bp. The PCR free library was quantified by Library Quantification Kit for Illumina Platforms (Roche Kapa Biosystems). The library was multiplexed with other libraries and the pool of libraries was then partly sequenced on a HiSeq4000 and partly on a HiSeq 2500 (TruSeq SBS Kit v4, Illumina, Inc) in paired-end mode (2 × 100 bp) in the Centro Nacional de Análisis Genómicos (CNAG, Barcelona). Primary data analysis, image analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (RTA 2.7.7) for HiSeq4000, and RTA1.18.66.3 when using HiSeq2500, followed by generation of FASTQ sequence files.

Additionally, four Loliinae samples (**Supplementary Table S1**) were used for Sanger sequencing of the nuclear ribosomal ITS locus and the plastid *trn*LF and *trn*TL loci using the primers and procedures indicated in Inda et al. (2008) in Macrogen and were added to the 97 Loliinae data set obtained from previous studies (Inda et al., 2008; Minaya et al., 2017).

DNA Sequence Data Assembling and Multiple Sequence Alignments

Illumina paired-end (PE) reads of the Fernandezian and other Loliinae samples were checked using FASTQC[2] and the adapters and low quality sequences were trimmed using TRIMMOMATIC (Bolger et al., 2014) at the CNAG. Plastome assembly was performed with Novoplasty v.2.7.1 (Dierckxsens et al., 2017) using the published plastomes of *Festuca ovina* (JX871940.1) for fine-leaved taxa and of *F. pratensis* (JX871941) for broad-leaved taxa (Hand et al., 2013) as reference, and the following parameters: k-mer: 27 or 39, insert size: ∼300 bp, genome range: 120,000–220,000 bp, and PE reads: 101 bp. Assembled plastomes were aligned using MAFFT v.7.031b (Katoh and Standley, 2013) followed by visual inspection using Geneious R11[3]. Because Novoplasty failed to assemble the whole plastome of *P. bromoides* due to the low number and quality of total PE reads, we used a Geneious mapping and readmerging strategy to map its reads to three phylogenetically close plastomes (*Megalachne berteroniana*, *M. masafuerana*, *Festuca pampeana*).

For the assembly of the nuclear ribosomal cistron we used a two-step read mapping and merging approach. Due to the lack of any published Loliinae rDNA cistron, we employed the *Brachypodium distachyon* rDNA cistron (reference genome Bd21, Vogel et al., 2010)[4] as reference and mapped to it the PE reads of the studied Loliinae taxa. Readmerging allowed us to align reads and their reverse complements to create a single consensus read. This step also allowed improving the sequence quality of overlapping parts. In cases of non-overlapping PE reads, the reads were used independently. The integrity of the cistron locus was examined visually for read mappings using Geneious R11.

Forward and reverse ITS, *trn*LF, and *trn*TL Sanger sequences were checked, corrected and merged using Sequencher v. 5.4.6 (Gene Codes Corporation, Ann Arbor, MI, United States)[5]. Each data set was aligned separately, visually inspected using Geneious R11 and manually corrected if necessary. The assembly of the *P. bromoides trn*LF and *trn*TL loci was done through several read mapping iterations with Geneious using as reference the closest *M. berteroniana*, *M. masafuerana*, *F. ventanicola* and *F. pampeana trn*LF and *trn*TL sequences.

A multiple sequence alignment (MSA) of 35 newly assembled *Megalachne* and Loliinae plastomes with *Oryza sativa* (AY522331.1; Genbank) and *Brachypodium distachyon* (NC_011032.1; Genbank) outgroups was performed with MAFFT v.7.215 (Katoh and Standley, 2013). The length of this full Loliinae plastome MSA without *Podophorus* was 146,172 bp length. The short *P. bromoides* consensus plastid sequence was subsequently aligned to the Loliinae full plastome MSA in Geneious R11. The multiple plastome alignment was filtered to remove poorly aligned regions and missing data in *P. bromoides* and other taxa through the automated option of trimAl v.1.2rev59 (Capella-Gutiérrez et al., 2009). The length of the filtered Loliinae plastome MSA with *Podophorus* was 55,872 bp length. A nuclear MSA of 35 newly assembled *Megalachne* and Loliinae rDNA cistrons and of *Oryza sativa* (AP008215; Genbank) and *Brachypodium distachyon* (Bd21; Phytozome) outgroups was also conducted with Geneious R11, rendering a 6,455 bp alignment. Independent MSAs were also produced for the ITS, *trn*LF, and *trn*TL loci of 135 Loliinae species and 16 outgroups, which included the *Megalachne* and *Podophorus* samples in Geneious R11. The *trn*LF and *trn*TL plastid loci were combined into a single plastid TLF MSA; separate phylogenetic analyses of the two loci gave congruent topologies with that recovered for the concatenated TLF haploid data matrix and only results from the latter analysis will be explained further. The nuclear ITS and the plastid TLF data set were further combined into a ITS-TLF MSA after obtaining congruence results from contrasted topological tests.

Phylogenetic Reconstruction and Divergence Time Analysis

Maximum likelihood phylogenetic analysis of the plastome (full and reduced), the rDNA cistron, and the independent and combined ITS, and TLF data sets were conducted with IQTREE (Nguyen et al., 2015) imposing the best-fit nucleotide substitution model to each separate data set that was automatically selected by the ModelFinder option of the program (Kalyaanamoorthy et al., 2017) according to the Bayesian Information Criterion (BIC) [plastome (full and reduced): TVM + F + R3; rDNA cistron: GTR + F + R2; ITS: SYM + I + G4; TLF: K3Pu + FR4]. Each search was performed through the automated computation of 20 Maximum Likelihood (ML) starting trees from 98 alternative

[2]http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3]https://www.geneious.com

[4]http://phytozome.jgi.doe.gov, v3.1Phytozome

[5]http://www.genecodes.com

randomized Maximum Parsimony (MP) trees, searching for best-scoring ML trees and estimating branch support for the best tree from 1,000 bootstrap replicates (BS) using the ultrafast bootstrap option (Minh et al., 2013; Chernomor et al., 2016) implemented in the software.

Ancestral divergence ages of the Fernandezian and other Loliinae grasses were estimated from the concatenated ITS-TLF data set with BEAST 2 (Bouckaert et al., 2014). We imposed independent site substitution models, lognormal relaxed clock and Yule tree models (Minaya et al., 2017). Two nodes of the Poaceae tree were calibrated using secondary age constrains for the crown nodes of the BOP clade (*Oryza* + Pooideae) (normal prior mean = 51.9 Ma, *SD* = 1.9) and the *Brachypodium* + core pooids clade (*Brachypodium* + Aveneae-Poeae) (normal prior mean = 30.9 Ma, *SD* = 3.5), following the grass-wide plastome based dating analysis of Sancho et al. (2018) and a third node was calibrated using a minimum age constrain (16 Ma) for the crown node of the fine-leaved Loliinae (lognormal prior mean = 19.5 Ma, *SD* = 0.101) based on a *Festuca* leaf macrofossil from Poland dated to the Early Miocene showing *Festuca* sect. *Festuca*-type adaxial and abaxial epidermises (Juchniewicz, 1975). We also imposed a broad uniform distribution prior for the uncorrelated lognormal distribution (ucld) mean (lower = 1.0E-6; upper = 0.1) and an exponential prior for ucld standard deviation (SD). We ran 600 million Markov chain Monte Carlo (MCMC) generations in BEAST2 with a sampling frequency of 1,000 generations. The adequacy of parameters was checked using TRACER v.1.6[6] with all the parameters showing Effective Sample Size (ESS) >200. A Maximum clade credibility (MCC) tree was computed after discarding 10% of the respective saved trees as burn-in.

Ancestral Range Estimation

We used the parametric dispersal-extinction-cladogenesis (DEC) approach implemented in Lagrange v. 20130526 (Ree and Smith, 2008) to infer global extinction and dispersal rates and ancestral range inheritance scenarios for each node representing the ancestors of the Fernandezian and other Loliinae grasses in the maximum clade credibility (MCC) tree obtained from BEAST. We defined 13 Operational Areas (OAs) (A-M), selected according to the current distribution ranges of the species and the potential historical distributions of their ancestors, delimited by geographical features that could have acted as barriers to dispersal (Minaya et al., 2017) (**Supplementary Table S2A**). Specifically, we selected four American OAs: North America (H), northern South America (I), southern South America (J), and Juan Fernandez (M), aiming to recover the areas of origin of the ancestors of *Megalachne* and *Podophorus* that presumably colonized the Juan Fernandez archipelago from the American mainland through long-distance dispersal (LDD). The ancestral ranges were built imposing a maximum of two ancestral areas (AA), considering that ancestors were not more widespread than their extant descendants (Sanmartín, 2003). Ancestral range inheritances and biogeographic events were inferred from a stratified model with four temporal windows (TSI: Late Oligocene to Middle Miocene, 28.4–16.0 Ma; TSII: Middle to Late Miocene, 16.1–7.2 Ma; TSIII: Late Miocene to Pliocene, 7.3–2.6 Ma; TSIV: Quaternary, 2.61–0 Ma). This model included the different temporal paleogeographical configurations of the Americas and other continents that might have affected the evolution and the distribution of the main Fernandezian and other Loliinae lineages (**Supplementary Table S2B**). In order to obtain a more detailed fine-scale reconstruction of the biogeographic events that resulted in the Fenandezian grasses, a second DEC analysis was performed for the lineages of the American-Vulpia-Pampas clade where the Fernandezian subclade was nested within (see section Results). This second analysis was performed using a pruned MCC dated subtree for the American-Vulpia-Pampas clade and five OAs representing the current and paleo-geographical distributions of the lineages (H-North-Central America, N-Pampas-Ventania, O-Andes, P-Masatierra, Q-Masafuera; **Supplementary Tables S2C,D**).

RESULTS

Loliinae Genome Sequence Data, Plastomes, and Nuclear rDNA Cistrons

Most of the studied Loliinae genome-skimming sequenced samples, including the newly studied *Festuca asplundii*, *F. caldasii*, *F. holubii*, *F. procera*, and *F. quadridentata*, yielded a large number of PE reads, ranging from 9,863 to 40,669 kbp (**Table 1** and **Supplementary Table S1**). The two *Megalachne* samples were below that threshold (*M. berteroniana* 5,288 kbp; *M. masafuerana* 6,134 kbp) but showed high quality reads. The 164 years old *Podophorus bromoides* type specimen sample rendered 6,694 kbp poor quality PE reads (**Table 1** and **Supplementary Table S1**).

Most plastid assemblies produced a single plastome contig with a deep coverage of >50x per sample that contained its two inverted repeat regions (IRa, IRb). However, Novoplasty assemblies of *Festuca durandoi, F. spectabilis, F. superba*, *F. molokaiensis*, *F. abyssinica*, and *Megalachne berteroniana* gave several small contigs and their full plastome assemblies were constructed with these contigs and the read mapping approach using Geneious and plastomes of their closest species as references. Plastome lengths of broad-leaved Loliinae ranged from 134,231 to 134,734 bp and those of fine-leaved Loliinae from 132,599 to 133,869 bp; these values agreed with the plastome lengths retrieved by Hand et al. (2013) for their two main Loliinae group taxa. The lengths of the *Megalachne berteroniana* (132,812 bp) and *M. masafuerana* (132,826 bp) plastomes fell within the fine-leaved Loliinae range. The PE reads of the newly assembled plastomes were deposited in GeneBank under BioProject PRJNA626668[7] with accessions numbers SAMN14647043–SAMN14647077 and SAMN14668162 (**Table 1** and **Supplementary Table S1**). The full Loliinae plastome MSA is available in Github[8]. The *Podophorus bromoides* plastid

[6]http://beast.bio.ed.ac.uk/Tracer

[7]https://www.ncbi.nlm.nih.gov/sra/PRJNA626668

[8]https://github.com/Bioflora/Loliinae_plastomes

consensus sequence (total length ∼69,238 bp) covered different non-overlapping fragments of the aligned Loliinae plastomes (∼40.7%) with a low coverage depth (10x to 1x). The plastid *P. bromoides* sequence (with its nucleotide positions mapped against the full Loliinae plastome MSA) is available in Github[9].

We obtained a single contig of 6,453–6,455 bp for the rDNA cistrons of the studied *Megalachne* and other Loliinae samples. Coverage depth was relatively constant across the rDNA cistron sequences in most cases (>10x). The newly sequenced rDNA cistrons were deposited in GeneBank with accessions numbers MT145276–MT145310 (**Table 1**). The low quality genomic sequence available in the DNA obtained from the *P. bromoides* specimen resulted in a low number of PE reads, which precluded the readmerging of its full rDNA cistron; however, it allowed the assembly of its entire ITS region (**Table 1** and **Supplementary Table S1**). The nuclear rDNA cistron of the studied *Megalachne* and other Loliinae grasses had a conserved structure along its transcriptional unit of 6–6.5 kb length, containing the 5'-ETS (724 bp), the 18S gene (1,818 bp), the ITS (585 bp), and the 25S gene (3,408 bp) regions of similar mean length to those of other grasses.

The nuclear ITS locus and the plastid *trn*LF and *trn*TL loci were filtered, respectively, from the assembled rDNA cistrons and plastomes for the *Megalachne* and Loliinae samples (**Table 1** and **Supplementary Table S1**). For *P. bromoides*, the complete ITS sequence was recovered with a coverage depth ranging from 10x to 1x and was deposited in Genbank under accession code MT022522 (**Supplementary Table S1**). Up to 60 and 70% of, respectively, the *P. bromoides trn*LF and *trn*TL sequences were recovered with a coverage depth of 10x (MSAs available in Github) (see footnote 9). The ITS and TLF sequences of the newly analyzed *F. andicola*, *F. longipes*, *F. vaginalis*, and *F. valdesii* were incorporated to the study and were deposited in Genbank under accession codes EF584922-EF592955-EF585009; KY368804-KY368856-KY368907; EF584977-EF584977-EF585111; MT022522-MT040974 – MT040975 (**Supplementary Table S1**).

Loliinae Plastome and Nuclear Phylogenomic Trees

The full plastome data set (two *Megalachne* and 33 additional Loliinae samples) included 133,894 filtered positions of which 7,480 were variable and 4,160 potentially informative. The best plastome ML phylogenetic tree (**Figure 2A** and **Supplementary Figure S1A**) recovered a fully resolved and highly supported topology with most branches having 100% bootstrap support (BS), and only three (94–99% BS) and one (77% BS) branches having strong to relatively good support. This Loliinae phylogenomic tree based on plastome data showed a main split of broad vs. fine-leaved Loliinae lineages, and successive splits within both the broad-leaved (Central-South American, Lojaconoa, Drymanthele/Tropical and South African, Leucopoa, Subbulbosae, Schedonorus) and the fine-leaved (American-Neozeylandic I, Eskia/American I, American-Vulpia-Pampas, Psilurus-Vulpia/Exaratae-Loretia (with intermediate Subulatae-Hawaiian nested within), Festuca, Aulaxyper, American II, Afroalpine) clades. *Megalachne berteroniana* and *M. masafuerana* plastome sequences formed a Fernandezian clade, sister to *F. pampeana* and nested within the southern American American-Vulpia-Pampas clade. Newly sequenced South American plastome samples fell within the fine-leaved American II [*F. fimbriata*, (*F. asplundii*, *F. procera*)] and American I [(*F. holubii*, *F. chimboracensis*)] clades, and within a Central-South American broad-leaved clade [(*F. caldasii*, (*F. superba*, (*F. quadridentata*, *F. amplissima*)))]. Fuegian *F. pyrogea* fell within the fine-leaved Festuca clade and the broad-leaved *F. fenas* clustered within the European Schedonorus clade (**Figure 2A** and **Supplementary Figure S1A**).

The reduced plastome data set, which included the *Podophorus* sample, had 55,872 positions of which 5,989 were variable and 823 potentially informative. The optimal ML tree (**Figure 2B** and **Supplementary Figure S1B**) recovered a topology that was also fully resolved and almost identical to that of the complete plastome data, though the branch support was slightly lower across the phylogenomic tree [all braches with full support except seven branches with strong (90–99%), three with good (70–89%), and one with weak (60%) BS]. In this phylogenomic tree, *P. bromoides* was resolved as sister to the *Megalachne* subclade (90% BS) and formed a fully supported Fernandezian clade, which was nested within the American-Vulpia-Pampas lineage (**Figure 2B** and **Supplementary Figure S1B**).

The nuclear rDNA cistron data set (two *Megalachne* and 33 additional Loliinae samples) included 6,455 positions of which 502 were variable and 321 potentially informative. The best ML tree (**Figure 3A** and **Supplementary Figure S1C**) retrieved a fully resolved topology; however, some internal branches were very short and showed very low support [21 branches with strong (90–99%), seven with good (70–89%), and seven with weak (60%) or very weak (<50%) BS]. The rDNA cistron-based phylogenetic tree showed the successive divergences of early diverging paraphyletic broad-leaved lineages (Tropical and South African, Drymanthele, Lojaconoa, Leucopoa, Central-South American, South-American, Schedonorus, Subbulbosae), which were in most cases poorly supported and included the intermediate Subulatae-Hawaiian nested within, and the more recent split of the strongly supported fine-leaved clade (97% BS). The topology of the fine-leaved group showed successive weakly to strongly supported lineage splits [(Eskia, ((Aulaxyper, Exaratae-Loretia, Festuca), (American-Vulpia-Pampas, Psilurus-Vulpia, Afroalpine, American-Neozeylandic I, American I, American II)))]. *Megalachne berteroniana* and *M. masafuerana* formed a fully supported Fernandezian clade based on the cistron sequences; this clade was close to other species of the American I (*F. holubii*, *F. chimborazensis*) and American II (*F. asplundii*, *F. fimbriata, F. procera*) assemblages, which together with the American-Neozeylandic I *F. gracillima* formed a well-supported fine-leaved South American clade (91% BS). *Festuca pyrogea* was reconstructed as sister to *F. ovina* within the strong Festuca clade. Within the broad-leaved lineages, the strongly supported Central-South American (*F. amplissima*, *F. quadridentata*) and (*F. superba*, *F. caldasii*) clades were resolved in different positions across the broad-leaved subtree, and *F. fenas* clustered

[9]https://github.com/Bioflora/Podophorus_plastome

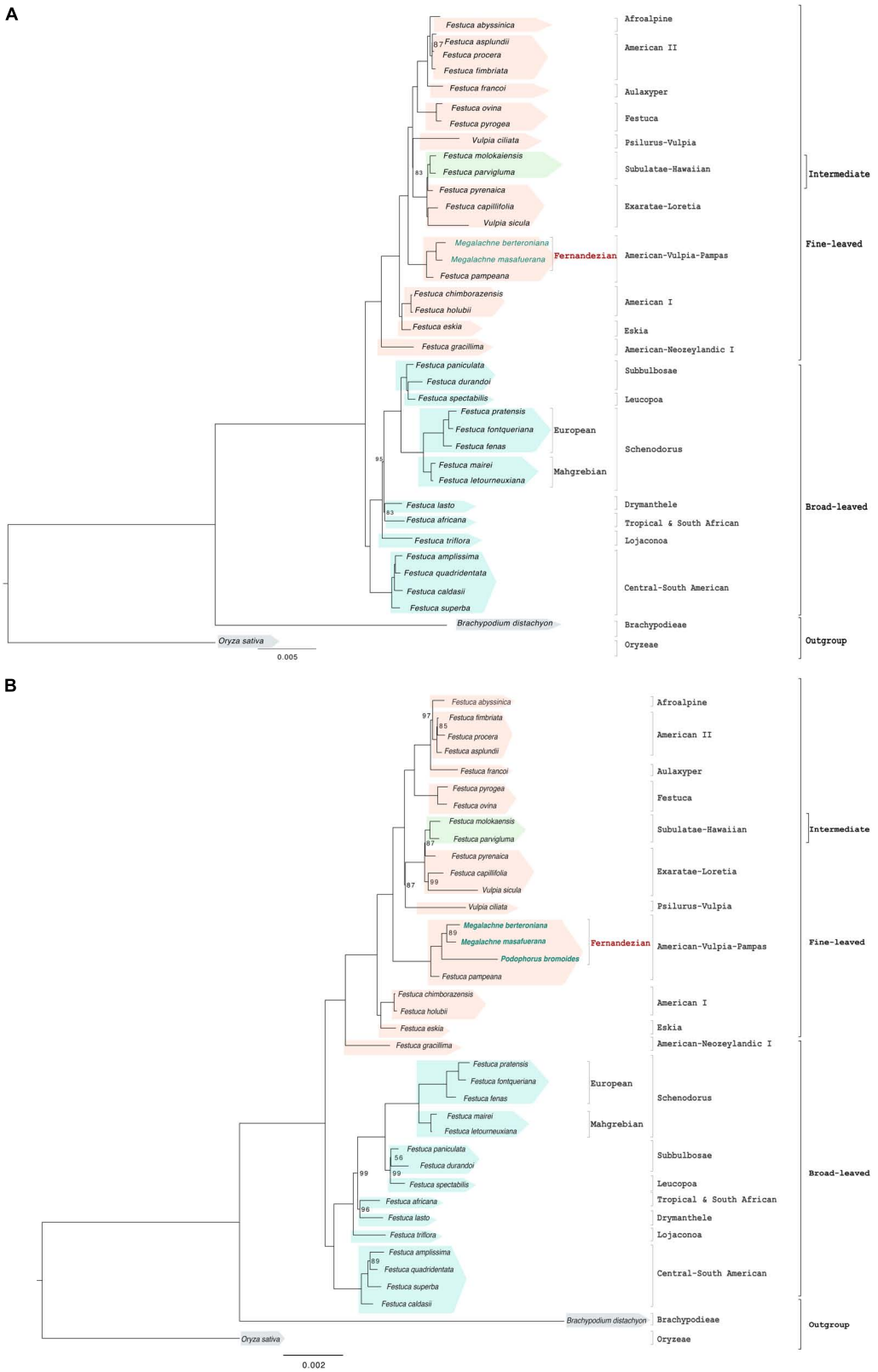

FIGURE 2 | Maximum likelihood full plastome **(A)** and reduced plastome **(B)** trees constructed with IQTREE showing the relationships among the studied Fernandezian and Loliinae grasses. *Oryza sativa* was used to root the trees. Numbers indicate branches with UltraFast Bootstrap supports (BS) <100%; the remaining branches have 100% BS value.

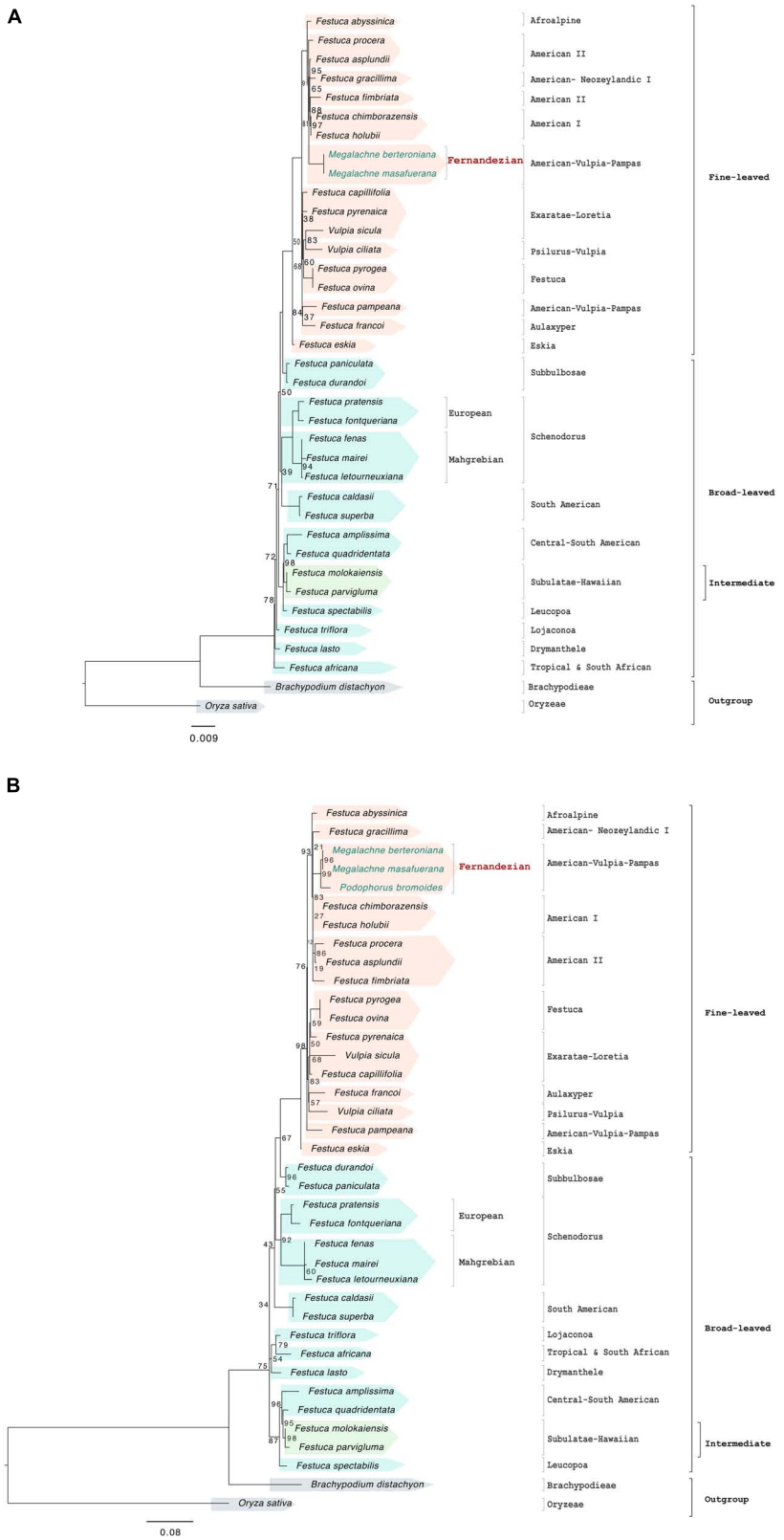

FIGURE 3 | Maximum likelihood nuclear rDNA cistron **(A)** and ITS **(B)** trees constructed with IQTREE showing the relationships among the studied Fernandezian and Loliinae grasses. *Oryza sativa* was used to root the trees. Numbers indicate branches with UltraFast Bootstrap supports (BS) <100%; the remaining branches have 100% BS value.

within the Mahgrebian Schedonorus subclade (**Figure 3A** and **Supplementary Figure S1C**). Phylogenetic reconstruction of filtered rDNA cistron sequences for the ITS region, together with that of *P. bromoides*, recovered the same overall tree topology, which showed a strong sister relationship of *P. bromoides* to the *Megalachne* clade (99% BS) (**Figure 3B** and **Supplementary Figure S1D**).

Plastid TLF, Nuclear ITS, and Combined ITS-TLF Phylogenetic Relationships

The separate and combined TLF (2,205 positions, 501 variable, 240 informative), ITS (645 positions, 285 variable, 193 informative) and ITS-TLF analyses of 135 Loliinae and outgroup samples retrieved phylogenies (**Supplementary Figures S2A–C**) highly congruent with those obtained in previous studies. Additionally, these trees showed the evolutionary placements of the three Fernandezian species and of six South American and one South African newly studied *Festuca* taxa. Both the nuclear ITS and the plastid TLF recovered a highly supported Fernandezian clade (99% BS) where *P. bromoides* was sister to the *M. berteroniana*/*M. masafuerana* subclade. Nonetheless, whereas the Fernandezian group was nested within a clade of American-Vulpia-Pampas taxa (69% BS), clearly separated from the American I (82% BS), and American II+Afroalpine (78% BS) clades in the TLF tree (**Supplementary Figure S2A**), it was nested within a large clade of American I + American II + Afroalpine taxa (99%) that also included some (*F. ventanicola*) but not all the American-Vulpia-Pampas species in the ITS tree (**Supplementary Figure S2B**). The combined ITS-TLF analysis placed the fully supported Fernandezian clade within a highly supported American-Vulpia-Pampas clade (97% BS) and resolved *F. ventanicola* as the strong sister lineage of the Fernandezian grasses (100% BS) (**Supplementary Figure S2C**). The TLF and ITS evolutionary placements of the newly sequenced South American taxa agreed with those of the plastome and rDNA trees and were overall congruent to each other. The fine-leaved *F. asplundii* and *F. procera* were nested within the American II + Afroalpine clade and *F. holubii* within the American I clade in the TLF tree (**Supplementary Figure S2A**), whereas the three of them fell within the large American I + American II + Afroalpine clade in the ITS tree (**Supplementary Figure S2B**). The sister *F. asplundii*/*F. andicola* (69% BS) and *F. holubii*/*F. glumosa* (87% BS) relationships observed in the ITS tree and their phylogenetic placements in the combined ITS-TLF tree (**Supplementary Figure S2C**) agreed with those of the plastid tree. The broad-leaved *F. quadridentata* and *F. caldasii* were nested within a large Central-South American-Eurasian-South African clade (97% BS) in the TLF tree (**Supplementary Figure S2A**) and in separate Central-South American (74% BS) and Eurasian-South American (62% BS) clades in the ITS tree (**Supplementary Figure S2B**). Their positions in the combined ITS-TLF tree (**Supplementary Figure S2C**) agreed with those of the nuclear tree. The South African *F. longipes* was resolved as sister of South African *F. scabra* (99% BS) in the TLF tree (Central-South American-Eurasian-South African clade) and of Tropical-South African *F. costata* in the ITS (100% BS) and combined ITS-TLF (88% BS) trees (Tropical-South African clade) (**Supplementary Figures S2A–C**).

Dating Analysis and Ancestral Range Inheritance Reconstruction

The Bayesian ITS-TLF MCC tree constructed with Beast2 (**Figure 4** and **Supplementary Figure S3**) yielded a similar topology to that retrieved in the ML analysis (**Supplementary Figure S2C**). The age of stem and crown Loliinae nodes were estimated to Late-Oligocene (median 21.47 Ma) and Early Miocene (19.4 Ma), respectively, whereas Early and Mid-Miocene divergences were inferred for the splits of the broad (16.31 Ma) and fine-leaved (16.83 Ma) lineages. An older Mid-Miocene origin was estimated for the ancestor of the American-Vulpia-Pampas clade (7.74 Ma) than for the younger Late-Miocene-to-Pliocene ancestors of the remaining fine-leaved [American II+Afroalpine (5.39 Ma); American I (3.89 Ma)] and broad-leaved [South-American (5.04 Ma); Central-South American (3.32 Ma)] South American Loliinae lineages (**Figure 4** and **Supplementary Figure S3**). The ancestor of the Fernandezian clade was inferred to have originated between the Late-Miocene (5.15 Ma; stem node) and the Pliocene (2.72 Ma; crown node), corresponding to the estimated split of *Podophorus* and *Megalachne*, whereas the split of the two *Megalachne* species was estimated to have occurred in the Pleistocene (1.02 Ma, Calabrian). The estimated ages of the Fernandezian ancestors predated those inferred for the ancestor of other oceanic endemic Loliinae lineages [e.g., Canarian fine-leaved Aulaxyper (4.11–2.84 Ma; Pliocene); Hawaiian *F. aloha*/*F. molokaiensis* (1.89–1.16 Ma; Lower-to-Recent Pleistocene); and recent Pleistocene Madeiran broad-leaved *F. donax* (1.23 Ma, Calabrian) and Reunion Island fine-leaved *F. borbonica* (0.3 Ma, Ionian)] (**Figure 4** and **Supplementary Figure S3**).

The ancestral range inheritance scenarios of Loliinae inferred from our Lagrange stratified Loliinae DEC model (-ln likelihood 404.6) had a global estimated dispersal rate (*dis*: 0.09385) 5.5 times higher than the estimated extinction rate (*ext*: 0.01536) (**Figure 5A**). The ancestors of Loliinae and of the broad-leaved and fine-leaved clades were inferred to have originated in uncertain widespread areas of the northern hemisphere (Mediterranean basin, Northern-Central America, Eurasia) in the transition between the Late Oligocene and the Early Miocene. Most of the transcontinental LDDs of both fine-leaved and broad-leaved Loliinae ancestors were estimated to have occurred during the Miocene and the Pliocene (time slices TSII-TSIII), and a few more during the Pleistocene (time slice TIV) (**Figure 5A**). According to our DEC model, the South American subcontinent was simultaneously colonized by broad and fine-leaved Mediterranean ancestors, which arrived, respectively, to the northern and southern South American ranges around the Mid-Miocene (**Figures 4**, **5A**). Within the fine-leaved lineage, a Mid-Miocene vicariance was inferred to have originated the American-Vulpia-Pampas ancestor in

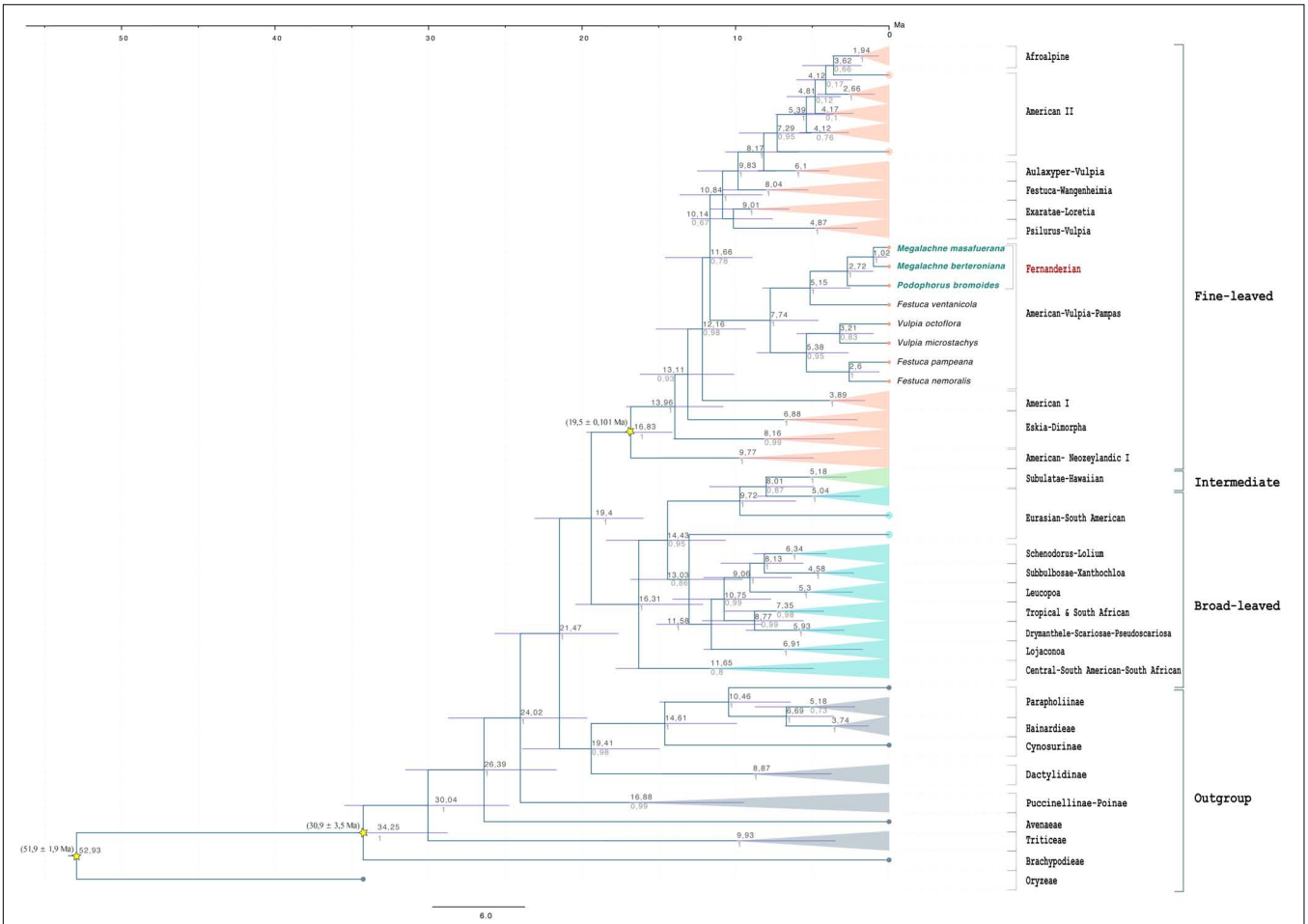

FIGURE 4 | Schematic Bayesian maximum clade credibility dated chronogram of 135 Loliinae taxa constructed with BEAST2 using nuclear ITS and plastid TLF loci showing estimated nodal divergence times (medians, in Ma) and 95% highest posterior density (HPD) intervals (bars) above branches and Posterior Probability Support (PPT) values below branches. Stars indicate secondary nodal calibration priors (means ± SD, in Mya) for the crown nodes of the BOP, *Brachypodium* + core pooids, and fine-leaved Loliinae clades.

southern South America ∼7.74 Ma. This ancestor would have then experienced range expansions to either North-Central America originating the southern American Pampean-Andean and the North-Central American Vulpia clade and to the Juan Fernandez archipelago originating the Pampean-Fernandezian clade at the end of the Neogene. Our stratified Loliinae DEC model suggested that the colonization of the Juan Fernandez archipelago from a mainland ancestor in southern South America could have occurred in the Mid-to-Late Miocene (7.74–5.15 Ma) (**Figures 4**, **5A**). According to this hypothesis, the ancestor of *F. ventanicola* and the Fernandezian *Podophorus* and *Megalachne* grasses was distributed in a widespread southern South America-Juan Fernandez area during the Late Miocene (5.15 Ma). A vicariance event was invoked to explain the split of the common ancestor into the current mainland Pampean-Ventanian endemic lineage and the Fernandezian ancestor, which was inferred to be present in the archipelago in the mid-Pliocene (2.72 Ma) (**Figures 4**, **5A**). A more detailed reconstruction of the biogeography of the Fernandezian grasses within their archipelago was obtained in our second American-Vulpia-Pampas DEC model (-ln likelihood 406.2; dis: 0.08232; ext: 0.0497) (**Figure 5B**). According to this model: (i) the ancestor of the American-Vulpia-Pampas could have been distributed in the Pampas-Ventanian range during the Miocene (7.74 Ma); (ii) this ancestor presumably experienced a range expansion to Masatierra and was present in a widespread Pampas-Fernandezian area during the Late Miocene (5.15 Ma); (iii) after a Pampas-Ventanian/Masatierra vicariance, the ancestor of the Fernandezian grasses was present in Masatierra during the Late Pliocene (2.72 Ma); (iv) an *in situ* speciation event originated the *Podophorus* lineage in Masatierra at that time; (v) a range expansion from Masatierra to Masafuera placed the ancestor of the *Megalachne* clade in the two main Juan Fernandez islands during the Pleistocene (1.02 Ma); (vi) a recent vicariance would explain the respective speciations of *M. berteroniana* in Masatierra and of *M. masafuerana* in Masafuera during the last million years (**Figures 4**, **5B**).

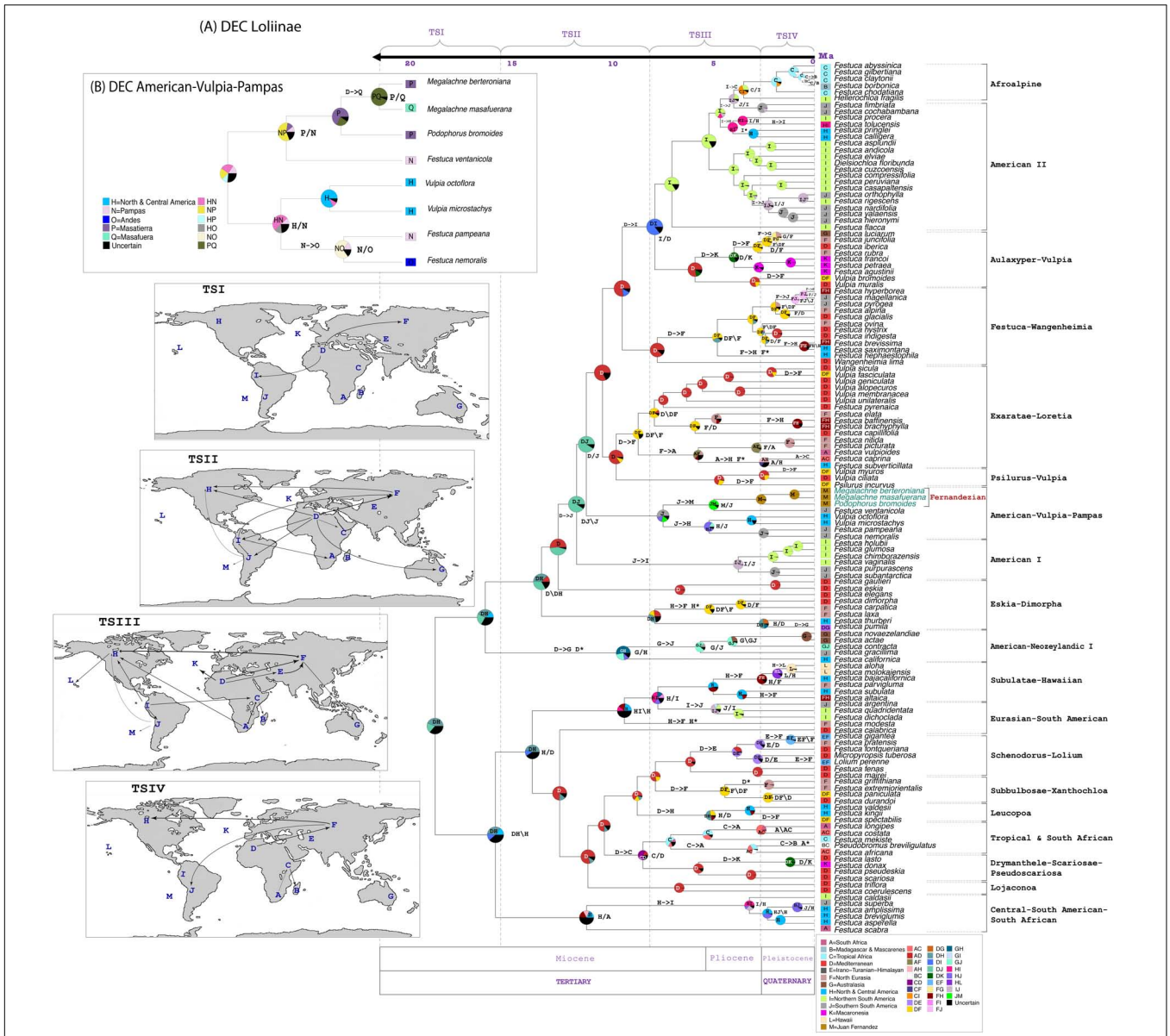

FIGURE 5 | Estimated ancestral ranges and biogeographical events of the Fernandezian and other Loliinae grasses inferred from LAGRANGE under the stratified DEC models mapped on the BEAST2 maximum clade credibility tree with outgroups pruned from it. **(A)** Loliinae DEC model; **(B)** American-Vulpia-Pampas DEC model. The vertical dashed lines separate the four time slices (TSI-TSIV) used in the Lagrange analyses. The maps on the left represent the palaeogeographical configuration of the world in these four time periods and the arrows represent the dispersals between areas that reflect change in continental connectivity over time. The pie charts at the nodes indicate the relative probabilities for alternative ancestral ranges (with their color legends indicated at the respective inset charts). The inferred biogeographic events are indicated at the nodes (X/Y vicariance; X\Y peripheral isolation) and branches (X->Y dispersal; X* extinction) of the tree. The Operational Areas assigned to the species are indicated to the right of the trees.

DISCUSSION

Phylogenetics of *Megalachne* and *Podophorus*: The Loliinae Fernandezian Clade

Our museomic approach, based on the combined use of old and recent herbarium samples and of genome skim data, have allowed us to disentangle the evolutionary origins of the neglected *Megalachne* and *Podophorus* grasses. Complete and partial plastomes as well as the nuclear rDNA cistron and ITS data supported the phylogenetic placement of the studied Fernandezian *Podophorous bromoides*, *Megalachne berteroniana*, and *M. masafuerana* species within the American-Vulpia-Pampas fine-leaved Loliinae clade (**Figures 2**, **3** and **Supplementary Figures S1A–C**). Our results corroborate the early suggestions of Tateoka (1962) who indicated a close

affinity of *Megalachne* to *Festuca* based on shared morphological and serological traits, and the recent phylogenetic findings of Schneider et al. (2011, 2012) and Tkach et al. (2020) who placed them within the fine-leaved Loliinae, and definitively discard its classification within either Bromeae or Duthieinae. Our results have also contributed to enlarge the paraphyly of *Festuca*, which now accounts to up to 14 Loliinae genera nested within its main fine-leaved (*Ctenopsis, Dielsiochloa, Hellerochloa, Megalachne, Micropyrum, Narduroides, Podophorus, Psilurus, Vulpia, Wangeheimia*), intermediate (*Castellia*), and broad-leaved (*Lolium, Micropyropsis, Pseudobromus*) lineages (**Supplementary Figure S2C**; Inda et al., 2008; Minaya et al., 2017).

Our study has demonstrated the utility of museomics to disentangle the evolutionary history of the extinct *Podophorus bromoides* from its 164 years old type specimen. This adds a new extinguished species to the tree-of-life, resolving its phylogenetic position within the grasses, as done before for other exterminated plants, such as *Sicyos villosus* within Cucurbitaceae (Sebastian et al., 2010) *Hesperelaea palmeri* within Oleaceae (Van de Paer et al., 2016; Zedane et al., 2016), *Haplostachys linearifolia* and *Stenogyne haliakalae* within Lamiaceae (Welch et al., 2016) and *Chasechloa egregia* within Poaceae (Silva et al., 2017). Moreover, our phylogenetic analyses based on plastome and rDNA-based data have demonstrated that *P. bromoides* is strongly resolved as sister to the *Megalachne* clade (*M. berteroniana*, *M. masafuerana*) (**Figures 2B**, **3B** and **Supplementary Figures S2A–C**), rejecting thus the moderately supported sister relationship found for the Masatierra taxa (i.e., *P. bromoides* and *M. berteroniana*, 72% BS) in a previous phylogenetic analysis based on partial ITS sequences from some samples (*Podophorus bromoides* ITS1 only) (Schneider et al., 2011).

Our Loliinae-wide phylogenomic analyses have further identified the relict Pampean-Ventanian fescues as the closest relatives of these fine-leaved endemic Fernandezian grasses. Phylogenies based on complete and partial plastome data indicate that *Megalachne* and *Podophorus* are strongly related to the American-Vulpia-Pampas lineage, represented by *F. pampeana* (**Figure 2** and **Supplementary Figure S2A**). By contrast, the nuclear rDNA cistron and the ITS phylogenies place them within a large American I + American II group (**Figure 3** and **Supplementary Figure S2B**), an assemblage that also includes other American-Vulpia-Pampas species, such as *F. ventanicola* (**Supplementary Figure S2B**). However, the phylogenetic tree reconstructed with the combined ITS-TLF data strongly supports nesting the Fernandezian clade within the American-Vulpia-Pampas clade and its sister relationship to the Pampean-Ventanian endemic *F. ventanicola* (**Figure 4** and **Supplementary Figures S2C, S3**). The incongruent placements of the Fernandezian grasses in the maternal plastome (plastid) vs. paternal rDNA cistron (ITS) Loliinae trees is a general feature of many Southern Hemisphere Loliinae species that reflect their hybrid allopolyploid nature (Inda et al., 2008; Minaya et al., 2017). Evolutionary studies have illustrated the different topological placements of known allopolyploid Loliinae species in plastid vs. nuclear trees (e.g., allotetraploid *F. fenas*, allohexaploid *F. arundinacea*, Inda et al., 2014; allotetraploid *F. simensis*, Inda et al., 2014; Minaya et al., 2015; allohexaploid *F. nigrescens*, Kergunteuil et al., 2020). Karyological and genome size reports have further shown that all southern hemisphere Loliinae species studied so far are polyploids (Dubcovsky and Martínez, 1992; Connor, 1998; Namaganda et al., 2006; Smarda and Stancik, 2006). Therefore, the incongruent positions shown by the American I clade polyploids *F. chimborazensis* (4*x*), *F. vaginalis* (4*x*), *F. glumosa* (4*x*), *F. purpurascens* (6*x*), American-Vulpia-Pampas clade *F. ventanicola* (4*x*) and the putative South African polyploid *F. longipes* in our plastid and nuclear trees (**Supplementary Table S1** and **Supplementary Figures S2A,B**) indicate that these taxa probably originated from interspecific hybridization followed by genome doubling. Although genome size or chromosome counting data are lacking for the Fernandezian *P. bromoides* and *M. berteroniana* and *M. masafuerana* species, their equivalent contrasting positions in the plastid and nuclear trees suggest that these endemic grasses are also allopolyploids. It is further supported by the fact that most of the remaining members of the American-Vulpia-Pampas clade are also polyploids [e.g., *F. pampeana* (8*x*), *F. nemoralis* (8*x*), *V. microstachys* (6*x*); Dubcovsky and Martínez, 1992; Smarda and Stancik, 2006; Díaz-Pérez et al., 2014]. Further investigation of these genomic data using the methodology described in Viruel et al. (2019) together with customized genome size analyses from fresh or herbarium samples (Smarda and Stancik, 2006) might reveal the ploidy level of these rare taxa.

Megalachne and *Podophorus* show a "vulpioid" phenotype, having lax panicles and long awned lemmas (**Figure 1**). These are characteristic traits of *Vulpia* and few other Loliinae lineages (Catalan et al., 2007). *Vulpia* and other ephemeral Loliinae genera, such as *Ctenopsis*, separate from *Festuca* based on their annual habit, four or less fertile florets per spikelet, largely unequal glumes, and long awned lemmas, which together distinguish them from the typical festucoid phenotype of *Festuca* and other robust Loliinae, characterized by their perennial habit, four or more fertile florets per spikelet, subequal glumes, and muticous or usually shortly awned lemmas, though none of them is absolute (Catalan et al., 2007). The origins of the polyphyletic *Vulpia* lineages are still intriguing although analysis of cloned single copy genes have demonstrated that some allopolyploid *Vulpia* species bear heterologous copies derived from morphologically close diploid relatives (Díaz-Pérez et al., 2014). The homoplasic "vulpioid" inflorescence phenotype has also appeared in other perennial Loliinae lineages, like the northern South America *Dielsiochloa floribunda* (American II clade), and in some species of *Festuca*. Interestingly, the slender cespitose Pampean-Ventanian endemic *F. ventanicola* shares its "vulpioid" phenotype with its sister Fernandezian *Megalachne* and *Podophorus* taxa (**Figure 1**), suggesting that they could have inherited it from their common ancestor. The long awn is an important dispersal trait in several annual grasses, including the invasive *Vulpia* species (Catalan et al., 2007; Díaz-Pérez et al., 2014), allowing the caryopsis to attach to the feathers or furs of animals and to be dispersed to long distances (Linder et al., 2018). It could be thus hypothetised that the

presumed "vulpioid" ancestor of the Fernandezian grasses could have migrated to the isolated Juan Fernandez archipelago transported by epizoochory or endozoochory through pelagic birds. Interestingly, *Podophorus bromoides* shows an extremely reduced spikelet (**Figure 1**), being the only Loliinae taxon, together with *Vulpia fontquerana* Melderis & Stace (Torrecilla et al., 2004) having a single fertile floret (with a reduced sterile floscule) per spikelet. This, together with its apparent ephemeral habit might be associated to an overall trend toward an annual habit after its speciation in the Masatierra island (**Figures 1**, **5**).

Biogeography and Conservation of the Endemic *Megalachne* and *Podophorus* Grasses

Our Loliinae and American-Vupia-Pampas biogeographic DEC analyses have elucidated the most likely colonization routes of the Fernandezian ancestors, and the speciation events that originated *Podophorus* and *Megalachne* taxa in Masatierra and Masafuera (**Figures 1**, **5**). Our ancestral range analyses identified the Pampean-Ventanian region to be the most likely place of origin for the common ancestors of the Fernandezian endemic grasses (**Figures 5A,B**). The closest relatives of *Podophorus* and *Megalachne* are relict endemic species of the Ventanian region (*F. ventanicola*, *F. pampeana*; Catalan and Müller, 2012) a hotspot of plant and animal diversity (Crisci et al., 2001). The formation of the Ventanian range in the Paleoproterozoic-Ordovician time span (∼2,200–475 Ma; Ramos et al., 2014) largely preceded the Oligocene-Miocene uplifting of the North American (31–28 Ma) and Central-Southern Andean (10–5 Ma) cordilleras (Crisci et al., 2001; Wakabayashi and Sawyer, 2001) as well as the emergence of the volcanic Juan Fernandez archipelago islands (5.8 Ma) (Stuessy et al., 1984). Although the inferred ages of the American-Vulpia-Pampas clade (7.7 Ma), *F. ventanicola* + Fernandezian clade (5.1 Ma) and Fernandezian clade (2.7 Ma) ancestors (**Figure 4** and **Supplementary Figure S3**) are younger than those of the Central-Southern Andes, the altitude and disposition of the austral Andean mountains was probably lower than in the present (Crisci et al., 2001). The geological time layout could have facilitated the hypothetical LDDs of the Ventanian ancestors to other American ranges and to the Juan Fernandez archipelago (**Figures 1**, **5**). Our Loliinae and American-Vulpia-Pampas DEC models support a colonization of the Fernandezian archipelago from a southern South American Pampean-Ventanian ancestor in the late-Miocene 7.7–5.1 Ma (**Figures 4**, **5**). The most recent estimate for that colonization concurs with the radiometric dating of the oldest Fernandezian islands (Santa Clara, 5.8 ± 2.1 Ma; Masatierra, 4.23 ± 0.16 Ma) (Stuessy et al., 1984) which could have been united in the past (Sanders et al., 1987). We could thus infer that the Ventanian Fernandezian ancestors likely arrived at the paleo-island formed by Santa Clara and Masatierra during the Late Miocene (**Figure 5**), probably transported by birds. The estimated split of the *Podophorus* lineage from the *Megalachne* ancestor at 2.7 Ma suggest a late-Pliocene *in situ* speciation event in Masatierra for the origin of the endemic *P. bromoides* (**Figure 4**). Our regional DEC model and our dating analyses infer that the colonization of the Masafuera island occurred from Masatierra during recent Pleistocene times (1.02 Ma), supporting *in situ* speciation events for *M. berteroniana* in Masatierra and *M. masafuerana* in Masafuera (**Figures 4**, **5B**). The westward inter-island colonization likely took place after the emergence of the young Masafuera island in the early Pleistocene (2.44 ± 1.14 Ma) (Stuessy et al., 1984) and was probably favored by the short distance separating them (i.e., 180 km, **Figure 1**). This distance has acted, however, as a strong geographic barrier to gene flow since the divergence of both species. Our biogeographic reconstruction for the Fernandezian Loliinae taxa agree with the hypothesis of higher levels of plant endemism in Masatierra compared to Masafuera, which are related to their respective distances to the closest mainland and their estimated ages (Stuessy et al., 2017). Our study has also identified the previously unknown South American ancestors of these endemic Fernandezian grasses, pointing to the relict Pampean-Ventanian region as their cradle (**Figure 5B**).

The rich endemic flora of Juan Fernandez archipelago is one of the most threatened on earth (Stuessy et al., 1998; Bernardello et al., 2006). Human impact on these islands, such as the introduction of environmentally aggressive herbivores, has probably caused the extinction of at least two endemic Fernandezian endemic plants during the last two centuries (*Santalum fernandezianum* Phil. and *Podophorus bromoides*; Bernardello et al., 2006; Danton et al., 2006). The latter extinct species was extremely rare; collected by Germain in 1854 and described by Philippi in 1856 from Masatierra (without a specific locotype), its existence was later mentioned by Johow in 1896 (Baeza et al., 2007). However, the plant was never seen again, even after exhaustive searches, and was therefore considered extinct (Stuessy et al., 1998, 2017; Baeza et al., 2007). All four *Megalachne* species are classified as threatened according to the IUCN categories of threat (Danton et al., 2006; Danton and Perrier, 2017; Penneckamp-Furniel, 2018; Penneckamp-Furniel and Villegas, 2019): *M. berteroniana* as Vulnerable, *M. masafuerana* as Endangered, *M. dantonii* as Critically Endangered, and *M. robinsoniana* as Endangered. Nonetheless, these IUCN assessments did not include a description of the employed IUCN criteria to classify the plants in their respective categories of menace. Several authors, however, have severe concerns about the threats posed to these endemic grasses by the introduced herbivores and by invasive plants (Stuessy et al., 1998; Bernardello et al., 2006; Danton et al., 2006; Danton and Perrier, 2017) and their survival in some inaccessible places to overgrazing pressure (Danton et al., 2006; Danton and Perrier, 2017). Rigorous populations censuses and population genetic studies of the more largely distributed *M. berteroniana* and *M. masafuerana* species, and of the recently described and still poorly known *M. robinsoniana* and *M. dantonii* species would be required to establish their adequate category of threat and to design appropriate conservation strategies. Historical collections have an enormous value for biogeographical studies. Several plants have gone to extinction in a few decades after human arrival due their high sensitivity to perturbation of their

habitats and their low competitiveness, especially in oceanic islands (Sebastian et al., 2010; Van de Paer et al., 2016; Welch et al., 2016; Zedane et al., 2016; Silva et al., 2017). Regrettably, *Podophorus bromoides* sums up to the list of recently extinct plants although its museomic analysis has unveiled its historical biogeography.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

PC designed the study. MM-A, IA, JV, and PC collected the samples. MM-A and JV developed the experimental work. PC, MM-A, JV, IA, and AS-R analyzed the data, interpreted the results, and revised the manuscript. PC and MM-A prepared the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Spanish Aragon Government and European Social Fund Bioflora A01-17R research grant. MM-A was supported by the University of Zaragoza-Santander Ph.D. fellowship.

ACKNOWLEDGMENTS

We thank Tod Stuessy for sending us herbarium samples of *Megalachne berteroniana* and *M. masafuerana*, the Kew Herbarium for facilitating the sampling of the *Podophorus bromoides* isotype (K000433684), the Ministerio del Ambiente of Ecuador for giving permission to collect Loliinae samples in the Ecuadorian paramos (MAE-DNB-CM-2015-0016), Antonio Diaz-Perez for assistance with the filtering of the *Brachypodium distachyon* rDNA cistron and three reviewers for their valuable comments to an early version of the manuscript. The genome skimming data of 35 Loliinae samples was generated at the Centro Nacional de Análisis Genómicos (CNAG, Barcelona, Spain) and that of *Podophorus bromoides* at Kew Botanical Gardens (United Kingdom). The bioinformatic and evolutionary analyses were performed at the Escuela Politécnica Superior de Huesca (Universidad de Zaragoza, Spain) Bioflora laboratory.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00819/full#supplementary-material

FIGURE S1 | (A) Maximum likelihood full plastome cladogram (35 Loliinae taxa, *Podophorus* excluded) constructed with IQTREE showing the relationships among the studied Fernandezian and Loliinae grasses. *Oryza sativa* was used to root the trees. Numbers indicate branches with UltraFast Bootstrap supports (BS). **(B)** Maximum likelihood reduced plastome cladogram (36 Loliinae taxa, *Podophorus* included) constructed with IQTREE showing the relationships among the studied Fernandezian and Loliinae grasses. *Oryza sativa* was used to root the trees. Numbers indicate branches with UltraFast Bootstrap supports (BS). **(C)** Maximum likelihood nuclear rDNA cistron cladogram (35 Loliinae taxa, *Podophorus* excluded) constructed with IQTREE showing the relationships among the studied Fernandezian and Loliinae grasses. *Oryza sativa* was used to root the trees. Numbers indicate branches with UltraFast Bootstrap supports (BS). **(D)** Maximum likelihood nuclear ITS cladogram (36 Loliinae taxa, *Podophorus* included) constructed with IQTREE showing the relationships among the studied Fernandezian and Loliinae grasses. *Oryza sativa* was used to root the trees. Numbers indicate branches with UltraFast Bootstrap supports (BS).

FIGURE S2 | (A) Maximum likelihood nuclear TLF tree (135 Loliinae taxa) constructed with IQTREE showing the relationships among the studied Fernandezian and Loliinae grasses. *Oryza sativa* was used to root the trees. Numbers indicate branches with UltraFast Bootstrap supports (BS) <100%; the remaining branches have 100% BS values. **(B)** Maximum likelihood plastid ITS tree (135 Loliinae taxa) constructed with IQTREE showing the relationships among the studied Fernandezian and Loliinae grasses. *Oryza sativa* was used to root the trees. Numbers indicate branches with UltraFast Bootstrap supports (BS) <100%; the remaining branches have 100% BS values. **(C)** Maximum likelihood combined ITS-TLF tree (135 Loliinae taxa) constructed with IQTREE showing the relationships among the studied Fernandezian and Loliinae grasses. *Oryza sativa* was used to root the trees. Numbers indicate branches with UltraFast Bootstrap supports (BS).

FIGURE S3 | Fully expanded Bayesian maximum clade credibility dated chronogram of 135 Loliinae taxa constructed with BEAST2 using nuclear ITS and plastid TLF loci showing estimated nodal divergence times (medians, in Ma) and 95% highest posterior density (HPD) intervals (bars) above branches and Posterior Probability Support (PPT) values below branches. Stars indicate secondary nodal calibration priors (means ± SD, in Mya) for the crown nodes of the BOP, *Brachypodium* + core pooids, and fine-leaved Loliinae clades.

TABLE S1 | List of taxa included in the phylogenetic study of the Fernandezian and other Loliinae grasses. Taxon, source, ploidy level, nuclear ITS, plastid *trn*TL and *trn*LF, plastome and rDNA cistron Genbank codes, average alignment insert size, total number of pair-end reads, number of plastome assembled reads, and number of rDNA cistron assembled reads are indicated for the corresponding samples.

TABLE S2 | (A) Operational areas used in the stratified Loliinae DEC Lagrange analysis. **(B)** Dispersal rate matrices reflecting the palaeogeographic connectivity among the study areas in each historical scenario (time slices TSI, TSII, TSIII, TIV). **(C)** Operational areas used in the stratified American-Vulpia-Pampas DEC Lagrange analysis. **(D)** Dispersal rate matrices reflecting the palaeogeographic connectivity among the study areas in each historical scenario (time slices TSIII, TIV).

REFERENCES

Baeza, C. M., Marticorena, C., Stuessy, T., Ruiz, E., and Negritto, M. (2007). Poaceae en el archipiélago de Juan Fernández (Robinson Crusoe). *Gayana Bot.* 64, 125–174.

Bernardello, G., Anderson, G. J., Stuessy, T. F., and Crawford, D. J. (2006). The angiosperm flora of the Archipelago Juan Fernandez (Chile): origin and dispersal. *Can. J. Bot.* 84, 1266–1281. doi: 10.1139/b06-092

Besnard, G., Christin, P.-A., Malé, P.-J. G., Lhuillier, E., Lauzeral, C., Coissac, E., et al. (2014). From museums to genomics: old herbarium specimens shed

light on a C3 to C4 transition. *J. Exp. Bot.* 65, 6711–6721. doi: 10.1093/jxb/eru395

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

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

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

Carter, K. A., Liston, A., Bassil, N. V., Alice, L. A., Bushakra, J. M., Sutherland, B. L., et al. (2020). Target capture sequencing unravels *Rubus* evolution. *Front. Plant Sci.* 10:1615. doi: 10.3389/fpls.2019.01615

Catalan, P. (2006). "Phylogeny and evolution of *Festuca* L. and related genera of subtribe Loliinae (Poeae, Poaceae)," in *Plant Genome: Biodiversity and Evolution*, eds A. K. Sharma and A. Sharma (Enfield: Science Publishers), 255–303.

Catalan, P., and Müller, J. (2012). "*Festuca* L," in *Flora de Argentina, Tomo II*, Vol. 3, eds A. M. Anton and F. O. Zuloaga (Buenos Aires: Estudio Sigma S.R.L), 219–250.

Catalan, P., Torrecilla, P., López-Rodríguez, J. A., Müller, J., and Stace, C. A. (2007). A systematic approach to subtribe Loliinae (Poaceae: Pooideae) based on phylogenetic evidence. *Aliso* 23, 380–405. doi: 10.5642/aliso.20072301.31

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

Connor, H. E. (1998). *Festuca* (Poeae: Gramineae) in New Zealand I. Indigenous taxa. *N. Z. J. Bot.* 36, 329–367. doi: 10.1080/0028825X.1998.9512574

Crisci, J., Freire-E, S., Sancho, G., and Katinas, L. (2001). Historical biogeography of the Asteraceae from Tandilia and Ventania mountain ranges (Buenos Aires, Argentina). *Caldasia* 23, 21–41.

Danton, P., and Perrier, C. (2017). Suppressions and additions to the flora of the Juan Fernández archipelago (Chile). *Bot. Lett.* 164, 351—-360. doi: 10.1080/23818107.2017.1396249

Danton, P., Perrier, C., and de Reyes, G. M. (2006). Nouveau catalogue de la flore vasculaire de L'archipel Juan Fernández (Chili) Nuevo catálogo de la flora vascular del Archipiélago Juan Fernández (Chile). *Acta Bot. Gallica* 153, 399–587. doi: 10.1080/12538078.2006.10515559

Diaz-Perez, A., López-Álvarez, D., Sancho, R., and Catalan, P. (2018). Reconstructing the origins and the biogeography of species' genomes in the highly reticulate allopolyploid-rich model grass genus *Brachypodium* using minimum evolution, coalescence and maximum likelihood approaches. *Mol. Phylogenet. Evol.* 127, 256–271. doi: 10.1016/j.ympev.2018.06.003

Díaz-Pérez, A. J., Sharifi-Tehrani, M., Inda, L. A., and Catalan, P. (2014). Polyphyly, gene-duplication and extensive allopolyploidy framed the evolution of the ephemeral *Vulpia* grasses and other fine-leaved Loliinae (Poaceae). *Mol. Phylogenet. Evol.* 79, 92–105. doi: 10.1016/j.ympev.2014.06.009

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45:e18. doi: 10.1093/nar/gkw955

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends Plant Sci.* 20, 525–527. doi: 10.1016/j.tplants.2015.06.012

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.

Dubcovsky, J., and Martínez, A. (1992). Distribución geográfica de los niveles de ploidía en *Festuca*. *Parodiana* 7, 91–99.

Greimler, J., López-Sepúlveda, P., and Reiter, K. (2017). "Chapter 6: vegetation," in *Plants of Oceanic Islands: Evolution, Biogeography, and Conservation of the Flora of the Juan Fernández (Robinson Crusoe) Archipelago*, eds T. Stuessy, D. Crawford, P. López-Sepúlveda, C. Baeza, and E. Ruiz (Cambridge: Cambridge University Press), 209–275.

Hackel, E. (1887). "Gramineae," in *Die Natürlichen Pflanzefamilien*, Vol. 2, eds A. Engler and K. Prantl (Leipzig: Verlag von Wilhelm Engelmann), 2–97.

Hand, M. L., Spangenberg, G. C., Forster, J. W., and Cogan, N. O. (2013). Plastome sequence determination and comparative analysis for members of the *Lolium-Festuca* grass species complex. *G3* 3, 607–616. doi: 10.1534/g3.112.005264

Harrison, N., and Kidner, C. A. (2011). Next–generation sequencing and systematics: what can a billion base pairs of DNA sequence data do for you? *Taxon* 60, 1552–1566. doi: 10.1002/tax.606002

Inda, L. A., Sanmartín, I., Buerki, S., and Catalan, P. (2014). Mediterranean origin and miocene–holocene old world diversification of meadow fescues and ryegrasses (*Festuca* subgenus *Schedonorus* and *Lolium*). *J. Biogeogr.* 41, 600–614. doi: 10.1111/jbi.12211

Inda, L. A., Segarra-Moragues, J. G., Müller, J., Peterson, P. M., and Catalan, P. (2008). Dated historical biogeography of the temperate Loliinae (Poaceae, Pooideae) grasses in the northern and southern hemispheres. *Mol. Phylogenet. Evol.* 46, 932–957. doi: 10.1016/j.ympev.2007.11.022

Juchniewicz, K. (1975). Flora kopalna Turowa koło Bogatyni w świetle analizy nabłonkowej. *Pr. Muzeum Ziemi* 24, 65–132.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., 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

Kellogg, E. A. (2015). "Flowering plants. Monocots. Poaceae," in *The Families and Genera of Vascular Plants*, Vol. XIII, ed. K. Kubitzki (New York, NY: Springer), 1–408.

Kergunteuil, A., Humair, L., Maire, A. L., Moreno-Aguilar, M. F., Godschalx, A., Catalan, P., et al. (2020). Tritrophic interactions follow phylogenetic escalation and climatic adaptation. *Sci. Rep.* 10:2074.

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.-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., Lehmann, C. E., Archibald, S., Osborne, C. P., and Richardson, D. M. (2018). Global grass (Poaceae) success underpinned by traits facilitating colonization, persistence and habitat transformation. *Biol. Rev. Camb. Philos. Soc.* 93, 1125–1144. doi: 10.1111/brv.12388

Malakasi, P., Bellot, S., Dee, R., and Grace, O. M. (2019). Museomics clarifies the classification of *Aloidendron* (Asphodelaceae), the iconic African tree aloes. *Front. Plant Sci.* 10:1227. doi: 10.3389/fpls.2019.01227

Minaya, M., Díaz-Pérez, A., Mason-Gamer, R., Pimentel, M., and Catalan, P. (2015). Evolution of the beta-amylase gene in the temperate grasses: non-purifying selection, recombination, semiparalogy, homeology and phylogenetic signal. *Mol. Phylogenet. Evol.* 91, 68–85. doi: 10.1016/j.ympev.2015.05.014

Minaya, M., Hackel, J., Namaganda, M., Brochmann, C., Vorontsova, M. S., Besnard, G., et al. (2017). Contrasting dispersal histories of broad-and fine-leaved temperate Loliinae grasses: range expansion, founder events, and the roles of distance and barriers. *J. Biogeogr.* 44, 1980–1993. doi: 10.1111/jbi.13012

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

Namaganda, M., Lye, K. A., Friebe, B., and Heun, M. (2006). AFLP-based differentiation of tropical African *Festuca* species compared to the European *Festuca* complex. *Theor. Appl. Genet.* 113, 1529–1538. doi: 10.1007/s00122-006-0400-5

Nevill, P. G., Zhong, X., Tonti-Filippini, J., Byrne, M., Hislop, M., Thiele, K., et al. (2020). Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. *Plant Methods* 16, 1–8.

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

Peña, C. M., Negritto, M. A., Ruiz, E., Baeza, C. M., and Finot, V. I. (2017). Revisión de *Megalachne* steud. (Poaceae: Pooideae: Poeae), género endémico

del Archipiélago de Juan Fernández, Chile. *Gayana Bot.* 74, 189–199. doi: 10.4067/S0717-66432017005000216

Penneckamp-Furniel, D. (2018). *Flora Vascular Silvestre del Archipiélago Juan Fernandez*, 1st Edn. Valparaíso: Planeta de papel ediciones.

Penneckamp-Furniel, D. P., and Villegas, G. R. (2019). A new species of *Megalachne* (Poaceae) endemic to Alejandro Selkirk Island, Juan Fernandez Archipelago, Chile. *Phytotaxa* 418, 294–300. doi: 10.11646/phytotaxa.418.3.5

Ramos, V. A., Chemale, F., Naipauer, M., and Pazos, P. J. (2014). A provenance study of the Paleozoic Ventania System (Argentina): transient complex sources from Western and Eastern Gondwana. *Gondwana Res.* 26, 719–740. doi: 10.1016/j.gr.2013.07.008

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

Richter, S., Schwarz, F., Hering, L., Böggemann, M., and Bleidorn, C. (2015). The utility of genome skimming for phylogenomic analyses as demonstrated for glycerid relationships (Annelida, Glyceridae). *Genome Biol. Evol.* 7, 3443–3462. doi: 10.1093/gbe/evv224

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

Sancho, R., Cantalapiedra, C. P., López-Alvarez, D., Gordon, S. P., Vogel, J. P., Catalan, P., et al. (2018). Comparative plastome genomics and phylogenomics of *Brachypodium*: flowering time signatures, introgression and recombination in recently diverged ecotypes. *New Phytol.* 218, 1631–1644. doi: 10.1111/nph.14926

Sanders, R. W., Stuessy, T. F., Marticorena, C., and Silva, O. (1987). Phytogeography and evolution of *Dendroseris* and *Robinsonia*, tree-Compositae of the Juan Fernandez Islands. *Opera Bot.* 92, 195–215.

Sanmartín, I. (2003). Dispersal vs. vicariance in the Mediterranean: historical biogeography of the Palearctic Pachydeminae (Coleoptera, Scarabaeoidea). *J. Biogeogr.* 30, 1883–1897. doi: 10.1046/j.0305-0270.2003.00982

Schneider, J., Winterfeld, G., Hoffmann, M. H., and Roeser, M. (2011). Duthieeae, a new tribe of grasses (Poaceae) identified among the early diverging lineages of subfamily Pooideae: molecular phylogenetics, morphological delineation, cytogenetics and biogeography. *Syst. Biodivers.* 9, 27–44. doi: 10.1080/14772000.2010.544339

Schneider, J., Winterfeld, G., and Roeser, M. (2012). Polyphyly of the grass tribe Hainardieae (Poaceae: Pooideae): identification of its different lineages based on molecular phylogenetics, including morphological and cytogenetic characteristics. *Org. Divers. Evol.* 12, 113–132. doi: 10.1007/s13127-012-0077-3

Sebastian, P., Schaefer, H., and Renner, S. S. (2010). Darwin's Galapagos gourd: providing new insights 175 years after his visit. *J. Biogeogr.* 37, 975–978. doi: 10.1111/j.1365-2699.2010.02270

Silva, C., Besnard, G., Piot, A., Razanatsoa, J., Oliveira, R. P., and Vorontsova, M. S. (2017). Museomics resolve the systematics of an endangered grass lineage endemic to north-western Madagascar. *Ann. Bot.* 119, 339—-351. doi: 10.1093/aob/mcw208

Smarda, P., and Stancik, D. (2006). Ploidy level variability in South American fescues (*Festuca* L., Poaceae): use of flow cytometry in up to 5 1/2-year-old caryopses and herbarium specimens. *Plant Biol.* 8, 73–80. doi: 10.1055/s-2005-872821

Soltis, D. E., Moore, M. J., Sessa, E. B., Smith, S. A., and Soltis, P. S. (2018). Using and navigating the plant tree of life. *Am. J. Bot.* 105, 287–290. doi: 10.1002/ajb2.1071

Soreng, R. J., Peterson, P. M., Davidse, G., Judziewicz, E. J., Zuloaga, F. O., Filgueiras, T. S., et al. (2003). Catalogue of new world grasses (Poaceae): IV. Subfamily Pooideae. *Contr. U.S. Natl. Herb.* 48, 1–730.

Soreng, R. J., Peterson, P. M., Romaschenko, K., Davidse, G., Teisher, J. K., Clark, L. G., et al. (2017). A worldwide phylogenetic classification of the Poaceae (Gramineae) II: an update and a comparison of two 2015 classifications. *J. Syst. Evol.* 55, 259–290. doi: 10.1111/jse.12262

Soreng, R. J., Peterson, P. M., Romaschenko, K., Davidse, G., Zuloaga, F. O., Judziewicz, E. J., et al. (2015). A worldwide phylogenetic classification of the Poaceae (Gramineae). *J. Syst. Evol.* 53, 117–137. doi: 10.1111/jse.12150

Spriggs, E. L., Eaton, D. A., Sweeney, P. W., Schlutius, C., Edwards, E. J., and Donoghue, M. J. (2019). Restriction-site-associated DNA sequencing reveals a cryptic *Viburnum* species on the North American coastal plain. *Syst. Biol.* 68, 187–203. doi: 10.1093/sysbio/syy084

Straub, S. C., 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

Stuessy, T., Crawford, D., and Ruiz, E. (2017). "Chapter 13: patterns of phylogeny," in *Plants of Oceanic Islands: Evolution, Biogeography, and Conservation of the Flora of the Juan Fernández (Robinson Crusoe) Archipelago*, eds T. Stuessy, D. Crawford, P. López-Sepúlveda, C. Baeza, and E. Ruiz (Cambridge: Cambridge University Press), 209–275.

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

Stuessy, T. F., Swenson, U., Crawford, D. J., Anderson, G., and Silva, O. (1998). Plant conservation in the Juan Fernandez archipelago, Chile. *Aliso* 16, 89–101. doi: 10.5642/aliso.19971602.04

Tateoka, T. (1962). Starch grains of the endosperm in grass systematics. *Bot. Mag.* 75, 377–383. doi: 10.15281/jplantres1887.75.377

Tkach, N., Schneider, J., Döring, E., Wölk, A., Hochbach, A., Nissen, J., et al. (2020). Phylogeny, morphology and the role of hybridization as driving force of evolution in grass tribes Aveneae and Poeae (Poaceae). *Taxon* doi: 10.1002/tax.12204

Torrecilla, P., López-Rodríguez, J.-A., and Catalan, P. (2004). Phylogenetic relationships of *Vulpia* and related genera (Poeae, Poaceae) based on analysis of ITS and trnL-F sequences. *Ann. Mo. Bot. Gard.* 91, 124–158.

Triantis, K., Whittaker, R. J., Fernández-Palacios, J. M., and Geist, D. J. (2016). Oceanic archipelagos: a perspective on the geodynamics and biogeography of the World's. *Front. Biogeogr.* 8:29605. doi: 10.21425/F5FBG29605

Van de Paer, C., Hong-Wa, C., Jeziorski, C., and Besnard, G. (2016). Mitogenomics of *Hesperelaea*, an extinct genus of Oleaceae. *Gene* 594, 197–202. doi: 10.1016/j.gene.2016.09.007

Viruel, J., Conejero, M., Hidalgo, O., Pokorny, L., Powell, R. F., Forest, F., et al. (2019). A target capture-based method to estimate ploidy from herbarium specimens. *Front. Plant Sci.* 10:937. doi: 10.3389/fpls.2019.00937

Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., et al. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463, 763–768. doi: 10.1038/nature08747

Wakabayashi, J., and Sawyer, T. L. (2001). Stream incision, tectonics, uplift, and evolution of topography of the Sierra Nevada, California. *J. Geol.* 109, 539–562. doi: 10.1086/321962

Welch, A. J., Collins, K., Ratan, A., Drautz-Moses, D. I., Schuster, S. C., and Lindqvist, C. (2016). The quest to resolve recent radiations: plastid phylogenomics of extinct and endangered Hawaiian endemic mints (Lamiaceae). *Mol. Phylogenet. Evol.* 99, 16–33. doi: 10.1016/j.ympev.2016.02.024

Zedane, L., Hong-Wa, C., Murienne, J., Jeziorski, C., Baldwin, B. G., and Besnard, G. (2016). Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea*, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. *Biol. J. Linn. Soc.* 117, 44–57. doi: 10.1111/bij.12509

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: 14 August 2020
doi: 10.3389/fpls.2020.01195

Unraveling the Spiraling Radiation: A Phylogenomic Analysis of Neotropical *Costus* L

*Eugenio Valderrama[1], Chodon Sass[2], Maria Pinilla-Vargas[1], David Skinner[3], Paul J. M. Maas[4], Hiltje Maas-van de Kamer[4], Jacob B. Landis[1], Clarice J. Guan[1] and Chelsea D. Specht[1]**

[1] School of Integrative Plant Science, Section of Plant Biology and the L.H. Bailey Hortorium, Cornell University, Ithaca, NY, United States, [2] The University and Jepson Herbaria, University of California, Berkeley, Berkeley, CA, United States, [3] Le Jardin Ombragé, Tallahassee, FL, United States, [4] Section Botany, Naturalis Biodiversity Center, Leiden, Netherlands

OPEN ACCESS

Edited by:
Lisa Pokorny,
National Institute of Agricultural and Food Research and Technology, Spain

Reviewed by:
Tomas Fer,
Charles University, Czechia
Roswitha Schmickl,
Academy of Sciences of the Czech Republic (ASCR), Czechia
Oriane Loiseau,
University of Lausanne, Switzerland

***Correspondence:**
Chelsea D. Specht
cdspecht@cornell.edu

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

Received: 06 February 2020
Accepted: 23 July 2020
Published: 14 August 2020

Citation:
Valderrama E, Sass C, Pinilla-Vargas M, Skinner D, Maas PJM, Maas-van de Kamer H, Landis JB, Guan CJ and Specht CD (2020) Unraveling the Spiraling Radiation: A Phylogenomic Analysis of Neotropical Costus L. Front. Plant Sci. 11:1195. doi: 10.3389/fpls.2020.01195

The family of pantropical spiral gingers (Costaceae Nakai; *c.* 125 spp.) can be used as a model to enhance our understanding of the mechanisms underlying Neotropical diversity. Costaceae has higher taxonomic diversity in South and Central America (*c.* 72 Neotropical species, *c.* 30 African, *c.* 23 Southeast Asian), particularly due to a radiation of Neotropical species of the genus *Costus* L. (*c.* 57 spp.). However, a well-supported phylogeny of the Neotropical spiral gingers including thorough sampling of proposed species encompassing their full morphologic and geographic variation is lacking, partly due to poor resolution recovered in previous analyses using a small sampling of loci. Here we use a phylogenomic approach to estimate the phylogeny of a sample of Neotropical *Costus* species using a targeted enrichment approach. Baits were designed to capture conserved elements' variable at the species level using available genomic sequences of *Costus* species and relatives. We obtained 832 loci (generating 791,954 aligned base pairs and 31,142 parsimony informative sites) for samples that encompassed the geographical and/or morphological diversity of some recognized species. Higher support values that improve the results of previous studies were obtained when including all the available loci, even those producing unresolved gene trees and having a low proportion of variable sites. Concatenation and coalescent-based species trees methods converge in almost the same topology suggesting a robust estimation of the relationships, even under the high levels of gene tree conflict presented here. The bait set design here presented made inferring a robust phylogeny to test taxonomic hypotheses possible and will improve our understanding of the origins of the charismatic diversity of the Neotropical spiral gingers.

Keywords: Costaceae, Zingiberales, plant radiation, phylogenomics, *Costus*, Neotropical

INTRODUCTION

One of the most widely recognized patterns in ecology and biogeography is that lineages tend toward species richness in tropical regions (Kreft and Jetz, 2007); however, the mechanisms that originate such patterns of diversity are still poorly understood. In addition, richness is not uniform across the tropical regions; the Neotropics stand as the most diverse with around 90,000–110,000 species of seed plants that could exceed the numbers of tropical Africa with 30,000–35,000 spp. and tropical Asia and Oceania with 40,000–82,000 spp., combined (Antonelli and Sanmartín, 2011; Hughes et al., 2013). Hypotheses addressing higher species richness in the Neotropics include opportunities for allopatric speciation, the availability of new habitats through uplift of the Andes (Gentry, 1982), major habitat and climate shifts prompted by shifts in the Amazon river drainage (Hoorn et al., 2010), and closure of the Panama isthmus (Bacon et al., 2013). Possibilities for prezygotic reproductive isolation driven by shifts in pollination syndromes (Serrano-Serrano et al., 2017), adaptation to local conditions leading to ecological speciation (Antonelli et al., 2018), or the effects of polyploidization on diversification rates (Soltis and Soltis, 2009; Landis et al., 2018) of Neotropical lineages are additional mechanisms that could explain the relatively higher diversity of Neotropical plant lineages compared to their Paleotropical congeners. Alternative explanations for the uneven distribution of biodiversity at continental scale include dispersal dynamics driven by historical changes in climate and differential extinction rates (Meseguer and Condamine, 2020). Specifically, the importance of extinction has been discussed to understand lower species richness in Africa compared to the Neotropics and South-East Asia (Couvreur, 2015).

The idea of the importance of interactions with pollinators for the diversification of flowering plants traces back to Darwin (1862). Selection can act to mold the characteristics of flowers driven by their predominant or most effective pollinators (Stebbins, 1970). The combination of traits (*e.g.* morphology, color, scent, size, rewards) associated with particular pollinator groups are known as pollination syndromes (Faegri and Pijl, 1979; Rosas-Guerrero et al., 2014). A recent study suggests that floral traits related to pollination efficiency (flower shape and orientation, position of reproductive organs) could be more important than widely considered traits including exposure, display size, scent, color, symmetry, and timing of anthesis (Dellinger et al., 2019). Although the validity of the concept of pollination syndromes has been debated, studies have been able to predict pollinators using floral traits and to confirm a stronger association in plants distributed in the tropics and associated with bats, bees, and hummingbirds (Rosas-Guerrero et al., 2014; Ashworth et al., 2015). Diversification rates within hummingbird pollinated lineages have been shown to be higher than in bee pollinated ones (Lagomarsino et al., 2016; Serrano-Serrano et al., 2017) and shifts towards hummingbird pollination syndrome associated with areas of high diversity of these birds in the Neotropics (Tripp and Manos, 2008). Furthermore, although syndromes can constitute specialized systems on specific pollinator guilds, they have been shown to be labile, with transitions and reversions happening repeatedly through the history of some Neotropical plant lineages (Tripp and Manos, 2008).

The family of pantropical spiral gingers (Costaceae Nakai; *c.* 125 spp.) can be used as a model to enhance our understanding of the mechanisms underlying Neotropical diversity. Costaceae has higher taxonomic diversity in South and Central America (*c.* 72 Neotropical species, *c.* 30 African, *c.* 23 Southeast Asian), particularly due to a radiation of Neotropical species of the genus *Costus* L. (*c.* 57 spp.). *Costus* is broadly distributed in the New World inhabiting lowland rain forest, montane rain forests, and periodically inundated várzea forests in elevations from the sea level up to 2,000 m, but mainly below 1,000 m (Maas, 1972). Previous studies have shown that the Neotropical species of *Costus* show multiple shifts in pollination syndromes, with closely related species that are associated with either insects or birds demonstrating rapid ecological isolation (Kay et al., 2005; Specht et al., 2012; Salzman et al., 2015). Furthermore, species within the Neotropical *Costus* clade have shown higher diversification rates during the last *c.* 10–20 million years (see André et al., 2016 for a discussion on the dates) as compared with the rest of the family, including the closely related African *Costus* lineages, and the prevalence in these lineages of sympatric species is higher regardless of time to differentiate (André et al., 2016). However, attempts to estimate phylogenies with a handful of plastid and nuclear loci have led to unresolved relationships in the species-rich clade comprising the Neotropical *Costus* (Salzman et al., 2015; André et al., 2016). Therefore, a well-supported phylogeny of the Neotropical spiral gingers, including thorough sampling of proposed species encompassing their full morphologic and geographic variation, is much needed.

The low resolution in the phylogenies adds uncertainty to the current understanding of the mechanisms that produced the charismatic and intriguing diversity within the spiral gingers. For example, a clear understanding of the phylogenetic relationships of closely related species that have undergone major shifts in morphology would allow us to test the genetic mechanisms underlying the changes between ornithophilous (bird attracting) and melittophilous (bee attracting) pollination syndromes that repeatedly took place in the history of this lineage and to characterize the role of these genetic mechanisms in shaping the speciation processes (Salzman et al., 2015). In addition, a fully resolved phylogeny of the species-rich clades of Costaceae would enlighten the taxonomy of the group (Maas, 1972; Maas, 1977), with extensive implications for understanding spatial and temporal patterns of distribution.

The difficulties in estimating robust, species-level phylogenies for speciose lineages are expected because of the combination of processes affecting recent radiations, including incomplete lineage sorting due to rapid differentiation and/or large population sizes and hybridization followed by introgression (Pamilo and Nei, 1988; Maddison, 1997; Maddison and Knowles, 2006). Coupled with the advances in sequencing technologies (Lemmon and Lemmon, 2013; McCormack and

Faircloth, 2013), target enrichment provides a solution for the need to acquire the hundreds or thousands of loci throughout the genome that are necessary to unveil the phylogenies of species rich and recently radiated plant lineages (Cronn et al., 2012). This is particularly true for those groups with large genome sizes, for which the sequencing and computational costs associated with whole-genome approaches quickly become restrictive as accession numbers increase (McKain et al., 2018). One of the additional and major advantages of targeted sequencing is that fragmented DNA from herbarium specimens can be used successfully (Hart et al., 2016; Brewer et al., 2019) allowing the sampling of lineages that are only available as herbarium specimens and to include specimens representing historic distributions. The accessions available for phylogenetic studies in natural history collections are essential to survey the diversity of species-rich groups, to include narrow endemics difficult to collect in the field and to account for variation in widespread and polymorphic species (Särkinen et al., 2012; Buerki and Baker, 2016; Bieker and Martin, 2018; Valderrama et al., 2018). The use of target enrichment strategies to gather low or single copy nuclear loci for phylogenomics of plant lineages at different scales (Nicholls et al., 2015; Sass et al., 2016) is becoming a standard technique, and the establishment of universal probe sets could reduce costs and time while enabling the merging of datasets from different studies and across plant lineages (Johnson et al., 2019; Larridon et al., 2020). However, divergence between the target sequences and the baits does affect capture efficiency (Larridon et al., 2020). The alternative process of designing custom baits allows researchers to aim for variable loci at the specific taxonomic scale of interest for the focus group, provided preliminary data is available for bait design (McKain et al., 2018). The increasing availability of genomic and transcriptomic data across the tree of life and the accessibility of pipelines to identify potential orthologs with low or single copy number (Chamala et al., 2015; Faircloth, 2016) help support the design of clade-specific bait sets (*e.g.* Vatanparast et al., 2018; Finch et al., 2019; Soto Gomez et al., 2019). Larridon et al. (2020) compared family specific probes and the Angiosperms-353 (Johnson et al., 2019) and obtained similar results with both approaches. However, universal probes could save labor and allow merging datasets of multiple studies, while taxa specific probes could improve recovery of target loci.

Here we use a phylogenomic approach to estimate the phylogeny of the Neotropical species of *Costus*, using a targeted enrichment approach. Baits were designed to capture conserved elements as identified from genomic sequences of *Costus* species and relatives. We sampled described and newly proposed species to test for reciprocal monophyly and included multiple samples from widespread and enigmatic species covering observed morphologic and geographic variation. DNA was extracted from living collections, field collected material, and herbarium samples to include population-level diversity. The resulting phylogeny of the Neotropical spiral gingers sheds light on the taxonomy of this lineage and enables us to confirm the multiple shifts in pollination syndromes during the evolution of *Costus* species.

MATERIALS AND METHODS

Taxon Sampling

Samples were chosen such that, when possible, they encompassed the geographical and/or morphological diversity of each species recognized or proposed for an updated monograph (Maas, 1972; Maas, 1977; Maas et al. pers. comm.). Widely distributed species or those being tested for monophyly include up to four accessions representing geographic and/or phenotypic variation. For field collected specimens, DNA was extracted from silica-dried leaf material, and voucher specimens were deposited in herbaria or in living collections (see **Supplementary Table 1**). For those not vouchered or in cultivation but included to increase geographic sampling for a given species, provenance data is recorded on inaturalist.org and is cross referenced with accession numbers. In total, thirty-one of *c.* 57 Neotropical *Costus* species were included in this analysis with sampling from field and herbarium-collected material.

Baits Design

Bait design followed the phyluce pipeline (Faircloth et al., 2012; Faircloth, 2016) with the following modifications. Instead of using annotated genomes and generating simulated reads from the assembled genomes, raw Illumina reads from *Costus spicatus* (Jacq.) Sw. and *Costus longibracteolatus* Maas genomic data (unpublished, Ana M.R. Almeida) were cleaned with TrimGalore 0.6.0 (Martin, 2011; https://github.com/FelixKrueger/TrimGalore) using a size cutoff of 36 bp (–length 36) and used in the alignment step. For the 7,723 regions that were found in the phyluce pipeline, local *de novo* assembly was performed with aTRAM 2.0 (Allen et al., 2018) using the cleaned *Costus* reads for each species separately, using two *de novo* assembly algorithms—Velvet 1.2.10 (Zerbino and Birney, 2008) and SPAdes 3.11.1 (Bankevich et al., 2012). Regions which generated a single *de novo* assembly contig after merging overlapping contigs (4-FinalAssembly.pl by Sonal Singhal; https://github.com/CGRL-QB3-UCBerkeley/denovoTargetCapturePopGen/blob/master/4-FinalAssembly) were carried on to subsequent filtering steps (2,686 regions). All regions that were found as a single contig in either *Costus* genome were carried forward; if the same region was found in both *Costus* genomes, the longer of the two regions was chosen. Several steps were added to the phyluce pipeline to filter regions of repetitive or putatively nonhomologous regions and to expand the dataset to regions that had known overlap with other published studies in the Zingiberales. 1) Sequences shorter than 160 bp were removed [2,388 regions remained]. 2) megaBLAST (Morgulis et al., 2008) all against all was conducted, and sequences which matched to any region other than itself were removed [removed 619 regions]. 3) BLAST (Altschul et al., 1990) searches against monocot mitochondrial and plastid genomes downloaded from the RefSeq database (O'Leary et al., 2016) were performed to remove sequences that matched these genomes [removed 399 regions]. At this point 2,019 regions passed filtering. 4) BLAST analyses to the RepeatMasker database (Smit et al., 2015) were used to identify

regions matching to transposons [removed five regions]. 5) Only regions with a GC content between 37 and 55% GC were retained to improve bait capture efficiency [removed two regions]. 6) Baits from a single *Costus* representative found in Sass et al. (2016) were added to the set [240 regions added]. 7) bait regions that were generated as part of Carlsen et al. (2018) were subjected to local *de novo* assembly with aTRAM as described above, to find these bait regions for *Costus* [47 regions added after filtering for length and GC content, as above]. Some regions were added that are of specific interest for studies addressing development and morphological characters (note: these were excluded from the downstream analyses of the present study) for a total target length of approximately 1 million base pairs. This dataset was used to create custom 100 mer probes in a 20 K design by myBaits (Arbor Biosciences, Ann Arbor, MI, USA) with 3× tiling.

DNA Extraction and Library Preparation

Leaf material was dried *in silica* and extracted using an SDS protocol (Edwards et al., 1991; Konieczny and Ausubel, 1993). Zymo DNA Clean & Concentrator-5 kits were used to purify the extractions (Zymo Research, Irvine, CA, USA). The size of the obtained fragments was checked in a 1% agarose gel. When average fragment size was above 350 bp, we followed the manufacturer's protocol for the Covaris E220 evolution Focused-ultrasonicator (Covaris, Woburn, MA, USA) to obtain an average fragment size of 350 bp. Double-sided-size selection was performed with size selection beads using a homemade solution of Carboxyl-modified Sera-Mag Magnetic Speed-beads (Thermo Fisher Scientific, Freemont, CA) in a PEG/NaCl buffer (Rowan et al., 2017).

Dual-indexed libraries were prepared following manufacturer's recommendations with the KAPA Hyper Prep kit with 500 ng of size-selected DNA quantified with Qubit 3.0 Fluorometer (Life Technologies, Grand Island, NY, USA). The volume per reaction was reduced to 1/5th following the recommendations of Lydia Smith at the Evolutionary Genetics Laboratory at UC Berkeley (comm. pers.; protocol available at https://osf.io/fkj2x). We used TruSeq style barcodes (8 bp) with a Stubby Adapter (see the **Supplementary Material Data**) and indexing primers provided by the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley. Indexed samples were pooled (4–10 samples/reaction) and enriched with the custom probes following the manufacturer's instructions (myBaits Manual v4.01, Arbor Biosciences, Ann Arbor, MI, USA) with a hybridization temperature of 65°C for 24 h. Because different blocking oligos show significant differences in performance (Portik et al., 2016), we used the Roche Universal Blocking Oligo Kit and SeqCap EZ Developer Reagent with plant C0t-1 DNA instead of the Blockers Mix supplied with the baits. Capture efficiency was assessed by comparing the amplification of target and off-target regions with a qPCR using the PowerUp™ SYBR™ Green Master Mix (Thermo Fisher Scientific Baltics UAB, Vilnius, Lithuania) in the ViiA 7 Real-Time PCR System (Applied Biosystems, Foster City, CA, USA). The enriched and pooled libraries (100 individuals in 11 reactions) were sequenced on a lane of NovaSeq SP 150PE in the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley.

Reads Processing, Assembly and Alignment

Reads were trimmed to remove low quality bases and adapter sequences with TrimGalore and normalized to 100× coverage using BBNorm (BBMap 38.74; Bushnell, 2020). HybPiper 1.3.1 (Johnson et al., 2016) with default settings was used to extract the reads that were mapped to the 1,521 target loci with BWA 0.7.12 (Li and Durbin, 2009). Mapped reads were assembled into contigs with SPAdes 3.13.1 (Bankevich et al., 2012) and discarded when coverage was lower than 8×. Summary statistics of the mapped reads were obtained with samtools 1.3 (Li et al., 2009). Only exonic sequences were kept in the downstream analyses to avoid inaccurate alignments. Paralog sequences for the assembled loci were retrieved with HybPiper. Loci with paralog warnings obtained for more than 5% of the accessions with recovered loci were excluded from downstream analyses. Available chloroplast genomes (Sass et al., 2016) were used to assemble plastid coding sequences using HybPiper and aTRAM; however, we recovered a very low amount of off-target reads in our libraries preventing us from generating comparable plastid sequences for our accessions. Contigs obtained were aligned using MAFFT 7.271 (Katoh and Toh, 2010) with the iterative (maximum iterations set to 10,000) refinement method incorporating local pairwise alignment information and with a gap opening penalty of 10. Trimal 1.3 (Capella-Gutiérrez et al., 2009) was used to remove poorly aligned bases and spurious sequences (-resoverlap and -seqoverlap parameters, 0.75. and 75 respectively).

Phylogenetic Inference

The alignments were used to estimate gene trees for each locus using RAxML 8.2.12 (Stamatakis, 2014) with the rapid bootstrap analysis (200 replicates) and search for best-scoring maximum likelihood tree in the same run with a GTR + GAMMA substitution model. Abnormally long branches were determined by TreeShrink (Mai and Mirarab, 2018) with default values for the species mode (α =0.05, b = 5%). The algorithm estimates the distribution of branch lengths for each individual within the gene trees and uses it to identify significantly long branches and removes them in the respective trees and alignments.

We concatenated the loci and fitted a GTR + GAMMA substitution model for each gene and allowed IQ-Tree 1.6.10 (Nguyen et al., 2015; Chernomor et al., 2016; Kalyaanamoorthy et al., 2017) to explore merging those partitions corresponding to each gene using the greedy heuristic algorithm (Lanfear et al., 2012) before finding trees. The analysis became computationally intractable when considering the many possible schemes to merge the partitions of so many genes. We therefore used the relaxed cluster algorithm (rcluster option; Lanfear et al., 2014)

that examines only the top 10% of the partition merging schemes. To assess the impact of using the relaxed cluster over the greedy heuristic algorithm, we also reduced the number of genes dividing the loci into three subsets to complete more thorough analyses using the greedy algorithm. Focusing on nodes with higher support within each gene tree (due to the overall low support values for individual gene trees), we used 40, 50, and 60% as threshold values of the upper quartile of rapid bootstrap support values obtained in RAxML for each gene tree to subset the obtained loci. This enabled us to focus on the loci that produced better supported trees and could potentially be more informative for our study.

We used ultrafast bootstrap approximation (Hoang et al., 2018) combined with the single branch SH-like approximate likelihood ratio test (SH-aLRT; Guindon et al., 2010) implemented in IQ-Tree, each with 10,000 replicates to assess the support of the resulting trees. The ultrafast bootstrap support values resulting from the analyses with the different subsets were mapped to the topology obtained with all loci using phangorn 2.5.5 (Schliep, 2011). Differences among subsets in ultrafast bootstrap support values were tested with a Friedman test (Friedman, 1937) and *post hoc* Wilcoxon signed-rank tests (Wilcoxon, 1945) with a Bonferroni correction (Bonferroni, 1935) in R 3.5.1 (R Core Team, 2013). Whenever possible, analyses were run in the CIPRES portal (Miller et al., 2011).

To consider incongruence among gene trees using methods statistically consistent under a multispecies coalescent model, we estimated species trees with ASTRAL 5.6.3 (Zhang et al., 2017) with all the obtained loci and the subsets. We contracted the low support branches of the gene trees (<10%) to improve the accuracy of the method (Zhang et al., 2017) using Newick Utilities 1.6 (Junier and Zdobnov, 2010). R packages treeio 1.10.0 and ggtree 2.0.4 (Yu et al., 2017) were used to plot the quartet support values estimated with ASTRAL on the resulting topology using the –t2 output option. We used phytools 0.6-99 (Revell, 2012) function cophylo to visually compare the concatenation and coalescent-based species trees.

Preliminary analysis indicated that the accessions from other Neotropical genera (*Dimerocostus* Kuntze and *Chamaecostus* C. Specht & D. W. Stev) were very divergent compared to the differentiation found within the Neotropical *Costus* lineages and could inflate the tree diameter and reduce the ability of TreeShrink to detect abnormally long branches, so only *Costus* species were included in the final analyses, with the African *C. fenestralis* Maas & H.Maas used as an outgroup based on previous studies confirming that Neotropical *Costus* are derived from African lineages (Salzman et al., 2015; André et al., 2016). Alignments with too few individuals (<50) and subsequently, individuals with too few loci (<520 for the analysis with all the obtained loci) were excluded from the analyses to avoid the effects of excessive missing data. Whenever necessary, accessions were removed from the alignments using AMAS 0.98 that was also used to generate summary statistics (Borowiec, 2016). The proportion of parsimony informative sites was compared among subsets with a Fisher–Pitman permutation test implemented in the R package coin 1.3-1 (Hothorn et al., 2008) using an approximative (Monte Carlo) reference distribution with 100,000 replicates and a *post hoc* pairwise permutation test with a Bonferroni correction to adjust *p* values for multiple comparisons with rcompanion 2.3.25 package (Mangiafico, 2016). Because of the assumed absence of hybridization and introgression transversal to the phylogenetic inference methods, all analyses were remade excluding the individuals identified as potential hybrids to avoid their impact on the results. The potential hybrids (nine individuals) and candidate parentals were identified based on morphological characters, and access to detailed images of those individuals is provided in **Supplementary Table 1**. We also estimated an evolutionary network for the New World *Costus* species using the NeighborNet algorithm with uncorrected p-distances and 500 bootstrap replicates in SplitsTree 4.16.1 (Huson and Bryant, 2005).

Phylogenetic Comparative Methods

To better understand the evolution of pollination syndromes in the Neotropical *Costus* clade we used stochastic character mapping (Huelsenbeck et al., 2003) to reconstruct ancestral character states. Taxa were coded as either bee pollinated (melittophilous) or bird pollinated (ornithophilous) based on their morphological display of pollination syndrome. We used models with equal and different transition rates for the shifts in pollination syndromes, as implemented in phytools, and generated 1,000 stochastic character maps with the resulting phylogeny of the concatenation approach. The equal and different rate models were compared with a likelihood-ratio test. Individuals of the same species that formed monophyletic clades were pruned from the phylogeny leaving a single accession per species. The resulting character maps were summarized to estimate posterior probabilities of the ancestral pollination syndromes of *Costus* diversity in the new world tropics. To explore biogeographical history of the study group, we assigned species to the World Wildlife Fund's ecoregions (Olson et al., 2001) as summarized by Antonelli et al. (2018). We used the data presented by Salzman et al. (2015) and from herbaria records available in the Global Biodiversity Information Facility to assign the areas to the species. Undescribed taxa and poorly known lineages were excluded to avoid underestimating the distribution ranges. Nonmonophyletic species were reduced to a single accession by keeping the one that matched the known phylogenetic affinities (Salzman et al., 2015; André et al., 2016). We used BioGeoBEARS likelihood framework to fit a model of Dispersal-Extinction Cladogenesis (DEC) to our dataset (Matzke, 2013), allowing any species to occupy a maximum of six areas of the eight included in the analysis. To fit a DEC model the tree was forced to be ultrametric using penalized likelihood with correlated rate variation among branches (Kim and Sanderson, 2008) using the chronos function of ape R package (Paradis and Schliep, 2019), and branch lengths were multiplied by 100,000 to have a range of values between 1 and 1,000. The +J model was not considered in the analysis because of its conceptual and statistical flaws (Ree and Sanmartín, 2018).

RESULTS

Capture Efficiency and Phylogenetic Information of Captured Reads

We obtained on average 4.018 (SD = 2.016, Min = 0.615–Max = 9.606) million reads per accession of which 46.612% (8.889%, 27.100–64.400%) were on target and assembled on average on 1,210.600 (248.501, 162–1,355) loci per accession (**Supplementary Figure 1**). Of the target loci intended for the phylogenomic reconstruction, we obtained 1,145 aligned loci generating 881,627 aligned base pairs yielding 36,596 parsimony informative sites (PIS). 313 loci had paralogy warnings for more than 5% of the obtained sequences; the remaining 832 had 792,974 aligned base pairs with 31,462 PIS. The distribution of loci that produced gene trees with higher bootstrap support values according to the thresholds (>40, >50, and >60%) of the upper quartile of the RAxML rapid bootstrap support values is presented in **Table 1** and **Supplementary Figure 2**. The longer alignments show a tendency to have more PIS (**Figure 1**), and the proportion of PIS is significantly different among the subsets of loci ($\chi^2[3] = 171$, $p < 0.0001$; **Figure 2**). The PIS are significantly higher in the subsets of loci that yielded the gene trees with at least 40% rapid bootstrap support values in the upper quartile (bs > 40% *v.* bs ≤ 40%: Z = 8.587, adjusted $p < 0.0001$; bs > 50% *v.* bs ≤ 40%: Z = 8.566, adjusted $p < 0.0001$; bs > 60% *v.* bs ≤ 40%: Z = 11.260, adjusted $p < 0.0001$) and marginally (bs > 50% *v.* bs > 60%: Z = −3.072, adjusted $p = 0.0128$) or nonsignificantly different among them (bs > 40% *v.* bs > 50%: Z = 0.794, adjusted $p > 0.999$; bs > 40% *v.* bs > 60%: Z = −1.838, adjusted $p = 0.396$).

TABLE 1 | Summary statistics of the length in base pairs and the number of parsimony informative sites (PIS) for the alignments of all the 832 loci and the subsets defined by the upper quartile of the RAxML rapid bootstrap support values of each gene tree (≤40, >40, >50, and >60%).

		Contig (bp)	PIS
All	**mean**	951.868	37.43
n = 832	**sd**	968.228	66.529
	min	126	0
	max	6,123	686
	total	791,954	31,142
≤40	**mean**	449.222	11.463
n =568	**sd**	411.547	21.41
	min	126	0
	max	3,515	290
	total	255,158	6,511
(40,50]	**mean**	1,684.18	69.876
n = 89	**sd**	957.432	77.98
	min	165	11
	max	5,895	595
	total	149,892	6,219
(50–60]	**mean**	2,031.622	87.418
n = 98	**sd**	847.812	90.043
	min	675	24
	max	6,123	686
	total	199,099	8,567
>60	**mean**	2,439.026	127.857
n = 77	**sd**	864.684	99.038
	min	487	45
	max	5,349	654
	total	187,805	9,845

Phylogenomic Inference

We obtained high support values for most of the inferred relationships using the concatenation approach (**Figure 3**). The ultrafast bootstrap support values obtained with the different subsets of loci are significantly different ($\chi^2[3] = 49.127$, $p < 0.0001$), and the analysis with the highest levels of support is the one that includes all available loci, as compared with analyses using only loci that produced more resolved gene trees and had a higher proportion of PIS (**Figure 4**). Wilcoxon signed-rank tests showed significant differences in the comparisons of the ultrafast bootstrap support values of All loci *v.* bs > 50% (V = 501, adjusted $p < 0.001$, adjusted $r = -0.454$), All loci *v.* bs > 60% (V = 622, adjusted $p < 0.0001$, adjusted $r = -0.540$), bs > 40% *v.* bs > 50% (V = 467.5, adjusted $p = 0.005$, adjusted $r = -0.365$) and bs > 40% *v.* bs > 60% (V = 612, adjusted $p < 0.0001$, adjusted $r = -0.518$). We obtained marginal differences for bs > 40% *v.* bs > 50% (V = 515.5, adjusted $p = 0.080$, adjusted $r = -0.228$) and nonsignificant differences for All loci *v.* bs > 40% (V = 273, adjusted $p = 0.607$, adjusted $r = -0.067$). All p values were corrected for multiple comparisons and subsequently used to estimate the r values. Considering a smaller subset of the best merging schemes of substitution models for the partitions did not prevent the analysis (including all loci) to yield higher support values. The topology remains stable when the number of regions included is reduced (except for the >60% subset), but support values decay when considering fewer loci, even if those being kept are the more informative ones within the dataset (**Supplementary Figures 3A–C**). The reduction in support values is most noteworthy for the deeper nodes in the tree comprising the early diverging lineages of Neotropical *Costus*. The branch lengths of the more weakly supported backbone of the phylogeny are very short, and the values of the local posterior probability of the ASTRAL analysis are also the lowest in the tree.

The normalized quartet score of the topology obtained with ASTRAL is 70.778%, suggesting high levels of discordance among gene trees. The quartet scores indicate high levels of gene tree conflict in the backbone of the phylogeny; even relationships with high local posterior probabilities show that several gene trees support the alternative topologies of each quartet (**Figure 5**). Despite the high levels of conflict among gene trees, short branches in the early diverging lineages of the phylogeny and the completely different approaches used to estimate species trees, the overall topology recovered with concatenation *v.* coalescent-based species tree method is almost identical, suggesting robustness of the relationships recovered by the methodology (**Figure 6**).

Most of the species which were sampled for more than one individual are recovered as monophyletic in our resulting phylogeny, even when considering broad geographical variation (*e.g. Costus lima* K. Schum. with individuals sampled from Ecuador and Costa Rica, *Costus lasius* Loes. with individuals from Peru and Panama) or morphological variation (*e.g. Costus* sp. nov. Peru with glabrous and pubescent forms recovered as sister). Enigmatic lineages that will likely constitute

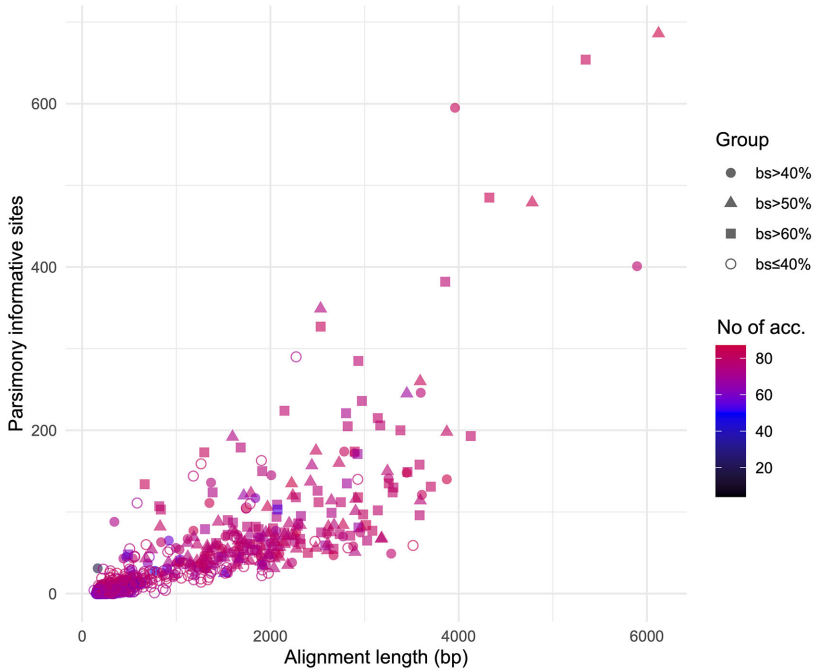

FIGURE 1 | Positive relation between alignment length and parsimony informative sites for the 832 loci obtained. Different shapes identify the subsets based on the threshold values of the upper quartile of rapid bootstrap support values obtained in RAxML for each gene tree. Colors indicate the number of accessions for which each loci was obtained.

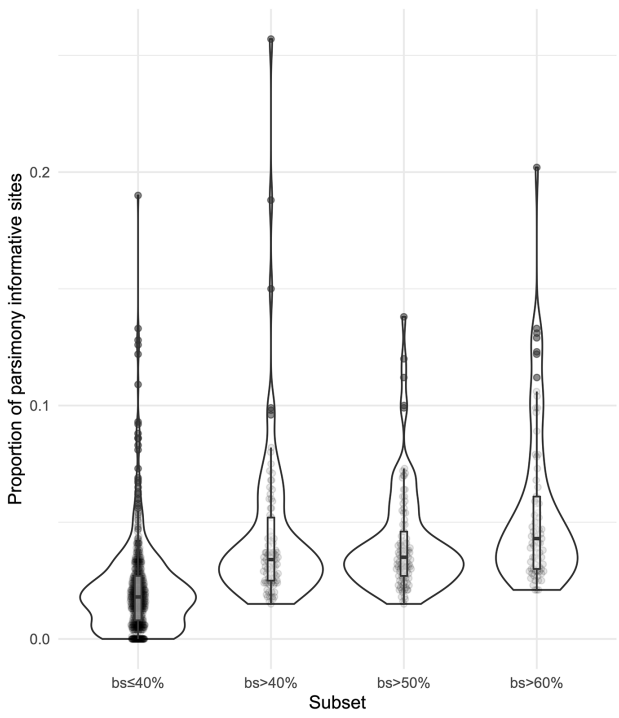

FIGURE 2 | Violin plots showing the distribution of the proportion of parsimony informative sites for the subsets of loci.

new species show considerable divergence from closely related species (*e.g. C.* sp. nov. Colombia). In other cases, our phylogeny includes lineages that are not closely related yet are currently considered as a single species: for example, *C. amazonicus* (Loes.) subspecies *amazonicus* J.F.Macbr. and *Costus amazonicus* subspecies *krukovii* Maas, and *C. guanaiensis* varieties (incl. *Costus guanaiensis* var. *tarmicus* (Loes.) Maas). Similarly, an individual from Puerto Rico identified as *Costus pictus* D.Don is not related to the accessions of the same species from Mexico and Costa Rica. Either *Costus aff. erythrothyrsus* accessions from the Acre Region in Brazil or *Costus erythrophyllus* Loes. lineages from the foothills of the eastern and western ridges of the Colombian Andes are monophyletic clades in our results. Various accessions having intermediate morphologies that were identified as potential hybrids between species cluster with one of the species identified as possible parentals. The support values for the backbone of the phylogeny are visibly lower in the analyses that included the potential hybrids (**Supplementary Figure 4**) than the analyses where those accessions were excluded (**Figure 3**). The NeighborNet network similarly clusters potential hybrids with candidate parentals and supports the topology obtained with the other analyses (**Supplementary Figure 5**).

Phylogenetic Comparative Methods

We selected the model with equal transition rates for the shifts in pollination syndromes for the stochastic character mapping analysis because including different rates did not

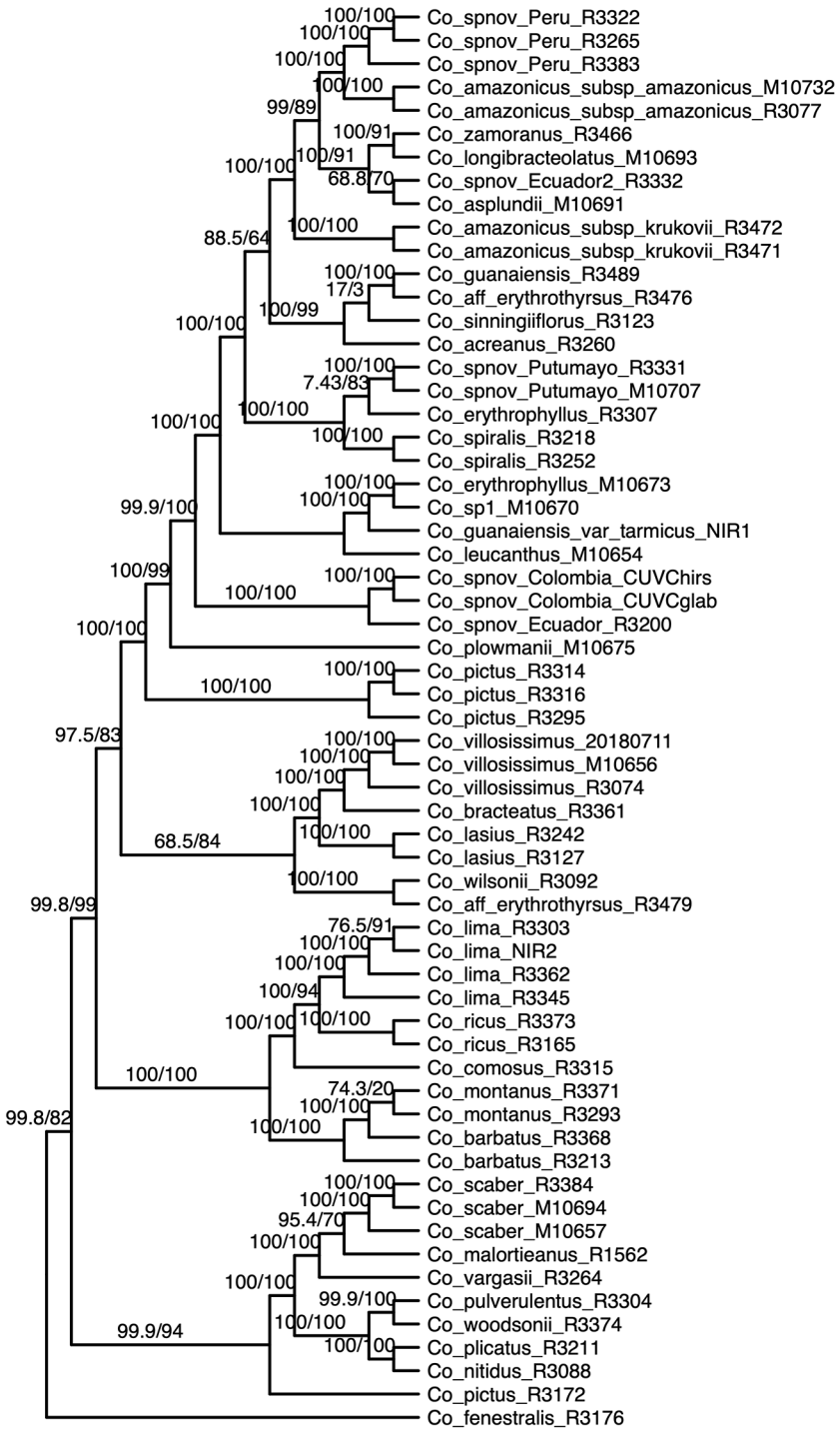

FIGURE 3 | Phylogenetic reconstruction with the concatenation of 832 loci analyzed in IQ-Tree; the values above the branches are the result of the SH-aLRT (above 80 are considered strongly supported) and ultrafast bootstrap support (above 95 are considered strongly supported) showing high support values in most of the branches. Equal branch lengths were used to allow the reader to distinguish support values; branch lengths are depicted in **Figure 6**.

improve likelihood significantly ($\chi^2[1] = 0.916$, $p = 0.339$). Posterior probabilities indicate multiple changes in pollination syndromes during the evolutionary history of *Costus*, with shifts occurring at least four times within the Neotropical lineage. The changes involve shifts to melittophilous pollination syndromes and subsequent regains of ornithophilous flowers. Our results suggest that the most recent common ancestor of all Neotropical *Costus* species was most likely ornithophilous

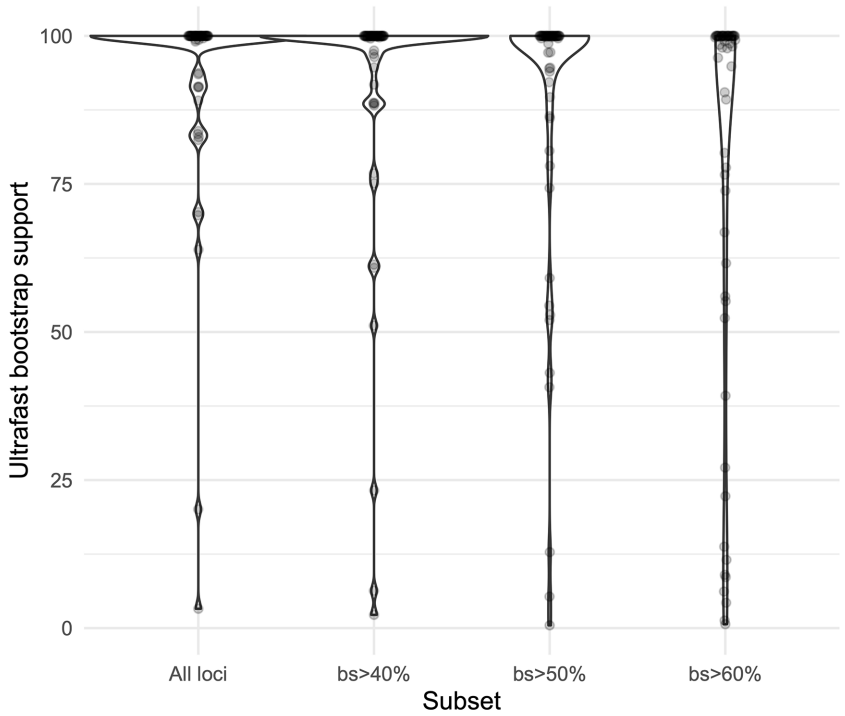

FIGURE 4 | Violin plots comparing the ultrafast bootstrap support values obtained with the concatenation of all the loci and the different subsets in IQ-Tree.

in form (**Figure 7**). The analysis reconstructing the evolution of the distribution range of *Costus* shows very high levels of uncertainty but also suggests a Central American origin for the genus (**Figure 8** and **Supplementary Figure 6**).

DISCUSSION

The custom-designed baits allowed us to gather informative loci for a good proportion of the sampled individuals. Phylogenetic signal recovered for the sampling of Neotropical *Costus* demonstrates the efficacy of using a targeted enrichment approach to estimate phylogenies in challenging plant lineages with large genomes, especially those involving rapid radiations, putative hybrids, and/or high levels of incomplete lineage sorting. The low proportion of reads recovered from the plastid genome prevented us from obtaining comparable sequences of the chloroplast and including them in the phylogenomic analysis. Our observed level of minimal capture of off-target reads has been documented in other studies (*e.g.* Villaverde et al., 2018; Forrest et al., 2019) and is perhaps attributable to highly efficient capture by our baits which were designed specifically for *Costus*. Studies that have particular interest in the plastid genome could still use similarly designed probes but increase the coverage of chloroplast regions by sequencing a mixture of captured and uncaptured libraries (Weitemier et al., 2014).

The phylogeny presented here considerably improves the resolution and support values of previous studies (Kay et al., 2005; Specht, 2006; Salzman et al., 2015; André et al., 2016), particularly providing resolution among the early branches (*i.e.* backbone) of the Neotropical *Costus* radiation. The branch lengths obtained along the backbone are relatively short, supporting the idea of a rapid radiation of the Neotropical lineages. Furthermore, normalized quartet score of the coalescent-based species tree topology indicates high levels of gene tree discordance, a result expected when incomplete lineage sorting is prevalent in the history of the group. Hybridization and the resulting introgression over the entire evolutionary history of the genus could also lead to the observed conflict in gene trees, contributing to the challenges in obtaining a well-supported phylogeny for the Neotropical *Costus*. Disentangling the influence of incomplete lineage sorting *v.* hybridization in our gene trees is not possible with the current sampling; however, more detailed sampling of various species complexes (*e.g. Costus comosus* (Jacq.) Roscoe; *Costus guanaiensis*) in the future could help detangle these processes particularly at the tips. Additional cases of nonmonophyletic species like *Costus amazonicus* and *Costus pictus* could be the pattern resulting from hybridization and introgression but also examples of cryptic species that require further studies on morphological and genomic evidence. Despite the challenging scenario of highly incongruent gene trees, the almost absolute concordance of the concatenation and coalescent-based species tree approach suggests that the topology obtained is stable, and the signal of the obtained loci overcomes the assumptions and caveats

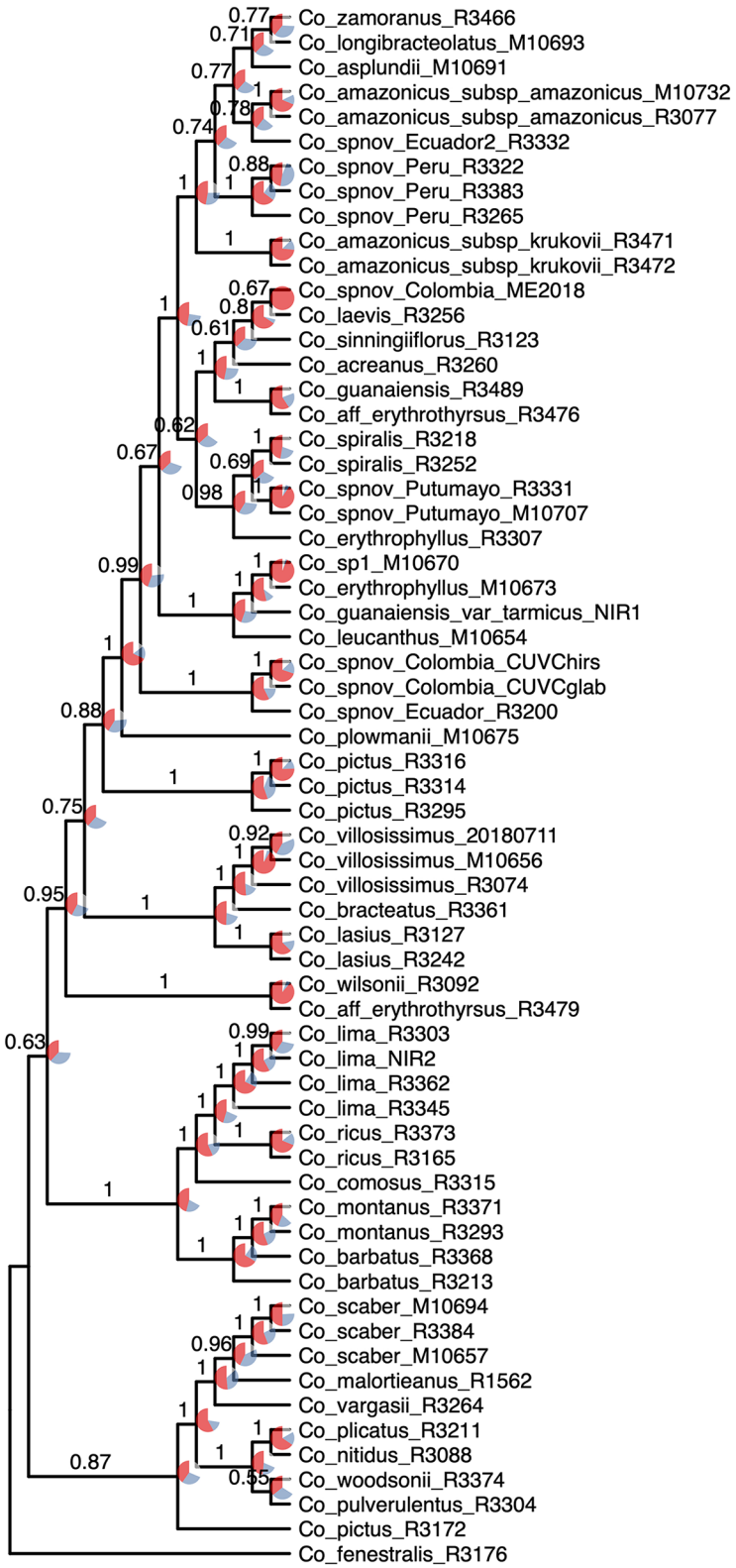

FIGURE 5 | Species tree reconstruction by ASTRAL with local posterior probabilities above the branches. Pie charts illustrate the quartet scores for each node for the 832 loci, with red representing the current topology, blue the second most favored topology, and white the remaining one. Equal branch lengths were used to allow the reader to distinguish support values; branch lengths are depicted in **Figure 6**.

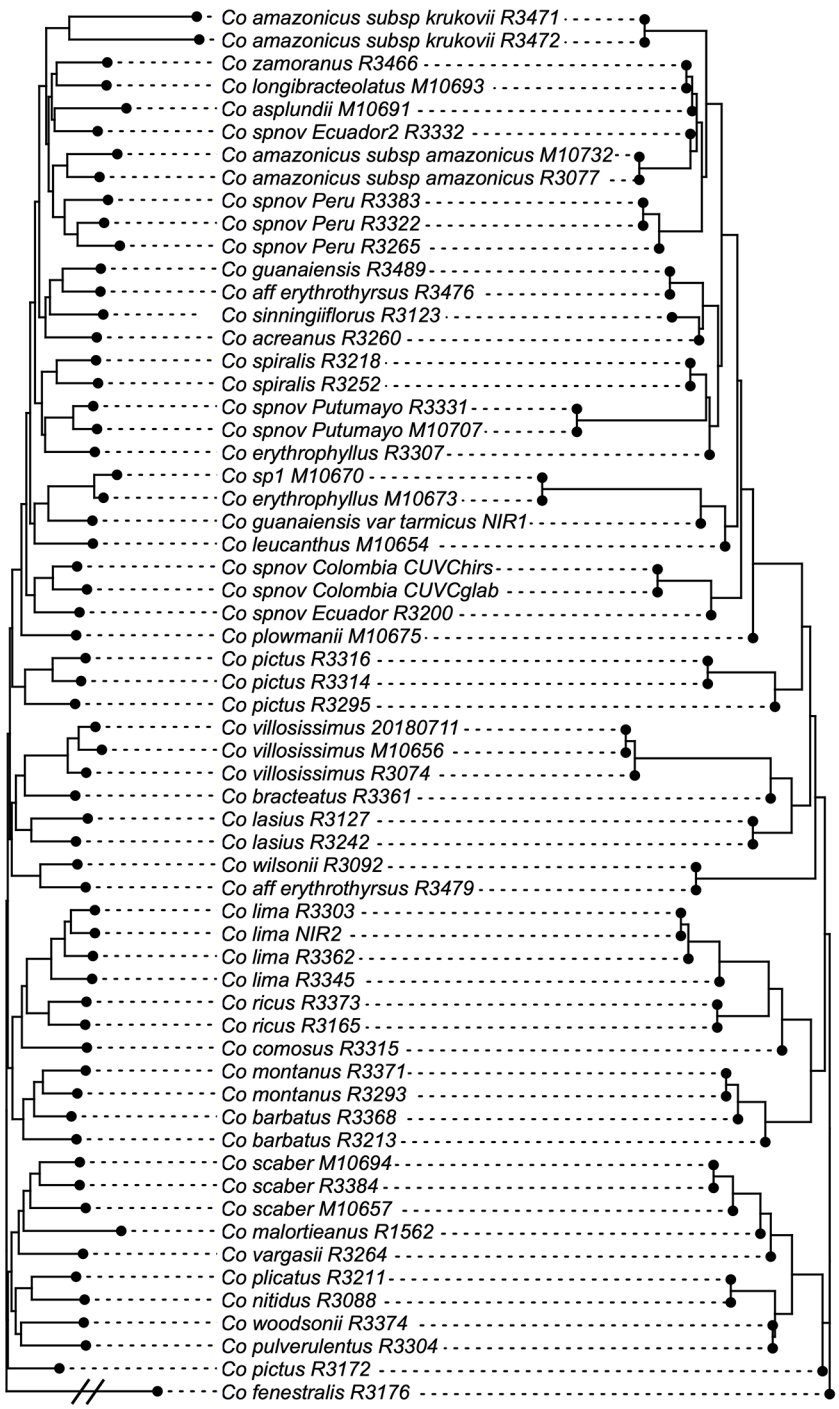

FIGURE 6 | Topologies obtained with concatenation *v.* coalescent-based species tree analyses, showing just one node difference between the two. Branch lengths proportional to the number of substitutions for the IQ-Tree result and to coalescent units in the ASTRAL result.

of the methods. The fact that the concatenation method produced the same topology as the method using a multispecies coalescent model, which explicitly accounts for incomplete lineage sorting, highlights the utility of concatenation-based methods for phylogenomic studies even in the presence of some degree of incomplete lineage sorting (Tonini et al., 2015; Streicher and Wiens, 2017). This is especially important given the high levels of gene tree incongruence present in this dataset.

Our observed decay in support values when building trees with reduced numbers of loci points to the importance of

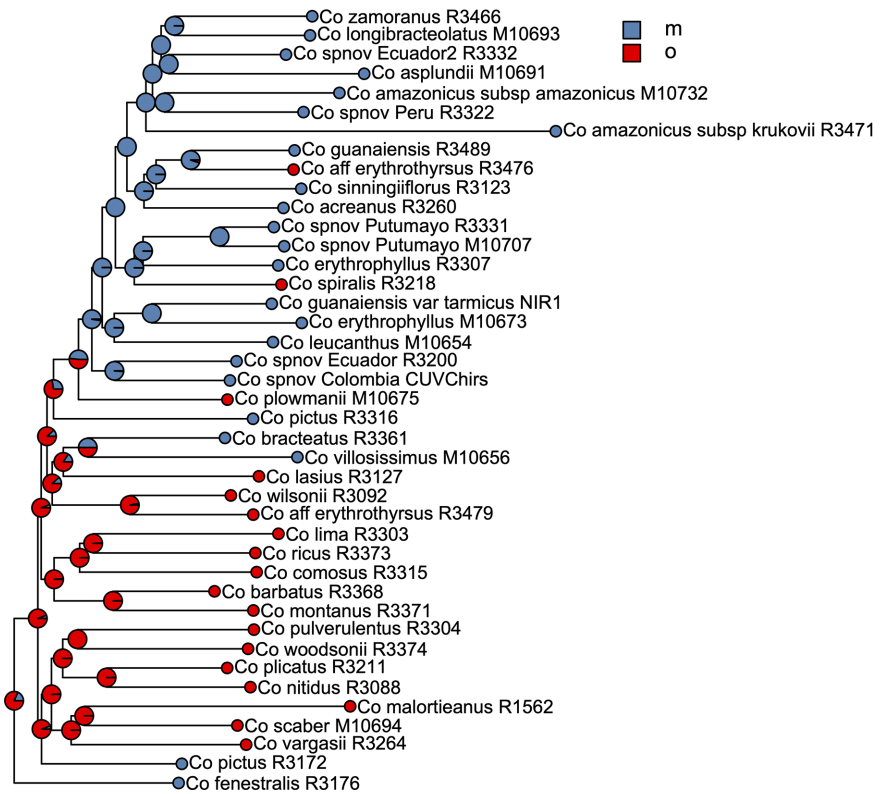

FIGURE 7 | Summary of the stochastic character mapping showing multiple shifts in pollination syndromes during the history of the Neotropical *Costus*. Pie charts indicate the posterior probabilities obtained from the 1,000 stochastic mappings (m, melittophilous; o, ornithophilous).

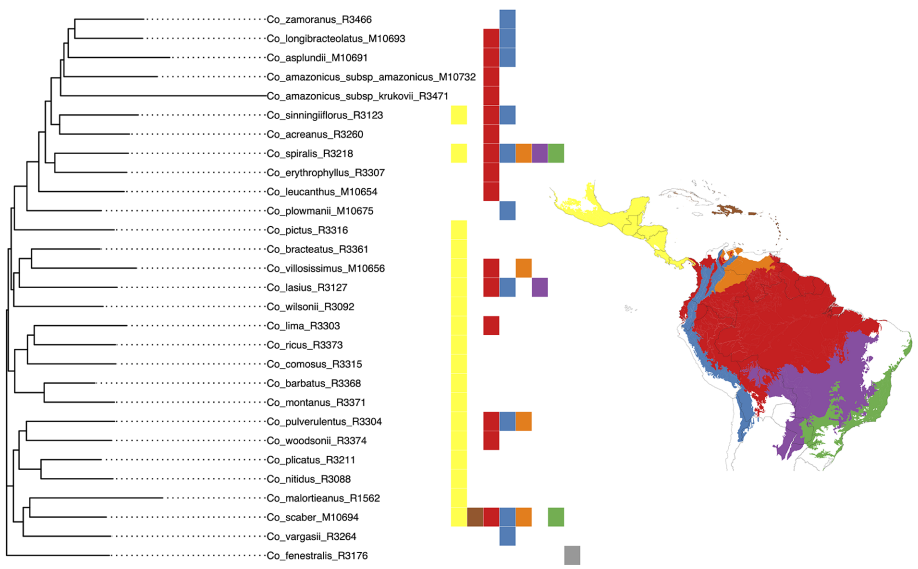

FIGURE 8 | Classification of the geographical distribution of the species of *Costus* included in the analyses. The regions from north to south are 1. Mesoamerica, 2. West Indies, 3. Amazon, Interandean Valleys and Choco-Darien region, 4. Northern and Central Andes, 5. Llanos region, 6. Cerrado, and 7. Atlantic Forest.

including as many loci as possible, ideally scattered across the genome (Blom et al., 2016; Bragg et al., 2018). The inclusion of more loci, even those with a lower proportion of parsimony informative sites and/or those generating poorly resolved gene trees, improved the support values of our resulting topology in concatenation analyses, particularly for the backbone where a lack of resolution has been emblematic for the Neotropical *Costus* clade. In our dataset, improvements in resolution obtained from including more loci overcome the computational restrictions in selecting schemes for merging partitions; this could be explained by the nonmutually exclusive effects of a very efficient solution for the heuristic problem (Lanfear et al., 2014) or the positive effect of gathering more phylogenetic signal when including more regions. It is important to highlight that the quartet scores indicate that relationships among the early diverging lineages of the Neotropical *Costus* show high discordance among gene trees. Even for some branches with relatively high local posterior probabilities, the quartet scores for the backbone of the current topology are low, suggesting that many loci support each of the alternative topologies in the quartets.

The ancestral area reconstruction shows very high uncertainty, probably due to the very short branches along the backbone of the phylogeny. Overall, our results agree with Salzman et al. (2015) in suggesting a Central American origin of Neotropical *Costus* species. Our results for the evolution of pollination syndrome morphology also agree with previous studies, indicating multiple shifts between bee- and bird-associated morphology occurring throughout the history of the genus. Results from stochastic character mapping suggest that the most recent common ancestor of all New World *Costus* most likely had a bird-pollinated form. Because most of the African species are insect pollinated (Maas-van de Kamer et al., 2016) and have either a melittophilous or generalist pollination form, our results point to an early appearance of the ornithophilous pollination syndrome in the ancestors of the Neotropical *Costus*. Furthermore, we confirm the reversal to a melittophilous form from ornithophilous morphology to have taken place at least twice and up to four times given our sampling (**Figure 7**). Interestingly, we also find evidence of regains of the bird pollinated flowers with high support in *Costus aff. erythrothyrsus* Loes. and *Costus spiralis* (Jacq.) Roscoe and with high uncertainty in *Costus plowmanii* Maas. These three lineages can be found at mid elevations (*c.* 1,000 m.), and the interaction with the highly diverse community of Neotropical montane birds (Quintero and Jetz, 2018) could have triggered those changes in morphology (Salzman et al., 2015). Establishing a temporal framework for these events will allow us to test the relationship of the shifts in pollination syndrome with the dramatic changes in the landscape that took place in the Neotropical region during the last 20 million years and elucidate the mechanisms that led to the high species richness in this clade perhaps resulting from an interaction between biotic and abiotic factors (Antonelli and Sanmartín, 2011). It is important to highlight that while including more species in our phylogeny and character mapping could change the specific results, overall agreement with the previous studies in the group suggests that the pattern of repeated shifts in overall floral form associated with pollinators is robust (Salzman et al., 2015; André et al., 2016).

Our phylogeny provides a guide for resolving problematic taxonomic hypothesis by testing and confirming monophyly when considering geographical and morphological variations within the described species. It also helps place enigmatic and undescribed lineages by comparing them carefully with their closest relatives. Some widely distributed and variable species are likely to be split into separate taxonomic units, thereby adjusting the taxonomy to accurately reflect evolutionary, morphological, and geographical variation. It is clear that diversity in the genus is underestimated by the current taxonomy and urges for an updated taxonomic revision. The potential to apply the baits described in this study to obtain similar datasets for a comprehensive sampling of all spiral gingers, including African taxa and the diversity only available as herbarium specimens, will allow us to test the hypothesis regarding the genetic mechanisms underlying the evolution of floral form and the recurrent changes in floral characters shown by closely related ornithophilous and melittophilous species. Finally, hybridization and introgression are likely to have been prevalent in the diversification of *Costus* in the Neotropics; a genome-wide dataset including comprehensive sampling of the diversity within the genus will allow us to test the prevalence and the directionality of hybridization events to better understand the role of reticulate evolution in the origin and diversification of the Neotropical spiral gingers.

DATA AVAILABILITY STATEMENT

The datasets and scripts generated for this study can be found in the Open Science Framework https://osf.io/fkj2x and raw reads in NCBI BioProject http://www.ncbi.nlm.nih.gov/bioproject/639561.

AUTHOR CONTRIBUTIONS

CDS conceived of the project and gathered the preliminary data. PM, HM-K, and DS provided cultivated and field-collected materials of otherwise impossible-to-get taxa representing documented morphologic and biogeographic variation. EV collected data, analyzed data, and wrote the manuscript. CDS, DS, PM, HM-K, and EV contributed to tissue collection, sampling and database management. MP-V and CG collected data and contributed to database management. CS collected and analyzed data. JL helped with analyses. All authors contributed to the article and approved the submitted version.

FUNDING

Research in this paper was supported by funds from Cornell University's College of Agriculture and Life Sciences and the

School of Integrative Plant Science. No federal support was used for this research.

ACKNOWLEDGMENTS

Authors are grateful to Susan Strickler and the Boyce Thompson Institute Computational Biology Center (BCBC) for providing access to computational resources. We thank Sidonie Bellot who provided valuable advice for plotting the figure with pie charts showing the quartet scores. Finally, we are grateful to Ana M. R. Almeida for providing unpublished data that was used for the baits.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.01195/full#supplementary-material

REFERENCES

Allen, J. M., LaFrance, R., Folk, R. A., Johnson, K. P., and Guralnick, R. P. (2018). aTRAM 2.0: An Improved, Flexible Locus Assembler for NGS Data. *Evol. Bioinforma.* 14, 1–4. doi: 10.1177/1176934318774546

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

André, T., Salzman, S., Wendt, T., and Specht, C. (2016). Speciation dynamics and biogeography of Neotropical spiral gingers (Costaceae). *Mol. Phylogenet. Evol.* 103, 55–63. doi: 10.1016/j.ympev.2016.07.008

Antonelli, A., and Sanmartín, I. (2011). Why are there so many plant species in the Neotropics? *Taxon* 60, 403–414. doi: 10.1002/tax.602010

Antonelli, A., Zizka, A., Carvalho, F. A., Scharn, R., Bacon, C. D., Silvestro, D., et al. (2018). Amazonia is the primary source of Neotropical biodiversity. *Proc. Natl. Acad. Sci. U. S. A.* 115, 6034–6039. doi: 10.1073/pnas.1713819115

Ashworth, L., Aguilar, R., Martén-Rodríguez, S., Lopezaraiza-Mikel, M., Avila-Sakar, G., Rosas-Guerrero, V., et al. (2015). "Pollination syndromes: A global pattern of convergent evolution driven by the most effective pollinator," in *Evolutionary Biology: Biodiversification from Genotype to Phenotype* (Switzerland: Springer International Publishing), 203–224. doi: 10.1007/978-3-319-19932-0_11

Bacon, C. D., Mora, A., Wagner, W. L., and Jaramillo, C. A. (2013). Testing geological models of evolution of the Isthmus of Panama in a phylogenetic framework. *Bot. J. Linn. Soc* 171, 287–300. doi: 10.1111/j.1095-8339.2012.01281.x

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

Bieker, V. C., and Martin, M. D. (2018). Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. *Bot. Lett.* 165, 409–418. doi: 10.1080/23818107.2018.1458651

Blom, M. P. K., Bragg, J. G., Potter, S., and Moritz, C. (2016). Accounting for Uncertainty in Gene Tree Estimation: Summary-Coalescent Species Tree Inference in a Challenging Radiation of Australian Lizards. *Syst. Biol.* 66, 352–366. doi: 10.1093/sysbio/syw089

Bonferroni, C. E. (1935). "Il calcolo delle assicurazioni su gruppi di teste," in *Scritti in onore di S. Ortu-Carboni* (Genoa, Italy: R. Istituto Superiore di Scienze economiche e commerciali).

Borowiec, M. L. (2016). AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 2016, 1–10. doi: 10.7717/peerj.1660

Bragg, J. G., Potter, S., Afonso Silva, A. C., Hoskin, C. J., Bai, B. Y. H., and Moritz, C. (2018). Phylogenomics of a rapid radiation: The Australian rainbow skinks. *BMC Evol. Biol.* 18, 1–12. doi: 10.1186/s12862-018-1130-4

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

Buerki, S., and Baker, W. J. (2016). Collections-based research in the genomic era. *Biol. J. Linn. Soc* 117, 5–10. doi: 10.1111/bij.12721

Bushnell, B. (2020). BBMap short read aligner, and other bioinformatic tools. Available at: https://sourceforge.net/projects/bbmap/ (Accessed January 22, 2020).

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinforma. Appl. Note* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

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

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. *Appl. Plant Sci.* 3, 1400115. doi: 10.3732/apps.1400115

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

Couvreur, T. L. P. (2015). Odd man out: why are there fewer plant species in African rain forests? *Plant Syst. Evol.* 301, 1299–1313. doi: 10.1007/s00606-014-1180-z

Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., et al. (2012). Targeted enrichment strategies for next-generation plant biology. *Am. J. Bot.* 99, 291–311. doi: 10.3732/ajb.1100356

Darwin, C. (1862). *On the various contrivances by which British and foreign orchids are fertilized* (London: Murray).

Dellinger, A. S., Chartier, M., Fernández-Fernández, D., Penneys, D. S., Alvear, M., Almeda, F., et al. (2019). Beyond buzz-pollination – departures from an adaptive plateau lead to new pollination syndromes. *New Phytol.* 221, 1136–1149. doi: 10.1111/nph.15468

Edwards, K., Johnstone, C., and Thompson, C. (1991). A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Res.* 19, 1349–1349. doi: 10.1093/nar/19.6.1349

Faegri, K., and Pijl, L. (1979). *The principles of pollination ecology. 3rd ed.* (Oxford: Pergamon Press).

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Syst. Biol.* 61, 717–726. doi: 10.1093/sysbio/sys004

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32, 786–788. doi: 10.1093/bioinformatics/btv646

Finch, K. N., Jones, F. A., and Cronn, R. C. (2019). Genomic resources for the Neotropical tree genus Cedrela (Meliaceae) and its relatives. *BMC Genomics* 20, 1–17. doi: 10.1186/s12864-018-5382-6

Forrest, L. L., Hart, M. L., Hughes, M., Wilson, H. P., Chung, K.-F., Tseng, Y.-H., et al. (2019). The Limits of Hyb-Seq for Herbarium Specimens: Impact of Preservation Techniques. *Front. Ecol. Evol.* 7, 439. doi: 10.3389/fevo.2019.00439

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* 32, 675–701. doi: 10.2307/2279372

Gentry, A. H. (1982). Neotropical floristic diversity: phytogeographical connections between Central and South America, Pleistocene climatic fluctuations, or an accident of the Andean orogeny? *Ann. Missouri Bot. Gard.* 69, 557–593. doi: 10.2307/2399084

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood

Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hart, M. L., Forrest, L. L., Nicholls, J. A., and Kidner, C. A. (2016). Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 65, 1081–1092. doi: 10.12705/655.9

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

Hoorn, C., Wesselingh, F. P., Ter Steege, H., Bermudez, M. A., Mora, A., Sevink, J., et al. (2010). Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science* 80 (330), 927–931. doi: 10.1126/science.1194585

Hothorn, T., Hornik, K., van de Wiel, M. A., and Zeileis, A. (2008). Implementing a class of permutation tests: the coin package. *J. Stat. Software* 28, 1–23. doi: 10.18637/jss.v028.i08

Huelsenbeck, J. P., Nielsen, R., and Bollback, J. P. (2003). Stochastic Mapping of Morphological Characters. *Syst. Biol.* 52, 131–158. doi: 10.1080/10635150390192780

Hughes, C. E., Pennington, R. T., and Antonelli, A. (2013). Neotropical Plant Evolution: Assembling the Big Picture. *Bot. J. Linn. Soc* 171, 1–18. doi: 10.1111/boj.12006

Huson, D. H., and Bryant, D. (2005). Application of Phylogenetic Networks in Evolutionary Studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

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

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

Katoh, K., and Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* 26, 1899–1900. doi: 10.1093/bioinformatics/btq224

Kay, K. M., Reeves, P. A., Olmstead, R. G., and Schemske, D. W. (2005). Rapid speciation and the evolution of hummingbird pollination in neotropical *Costus* subgenus *Costus* (Costaceae): Evidence from nrDNA its and ETS sequences. *Am. J. Bot.* 92, 1899–1910. doi: 10.3732/ajb.92.11.1899

Kim, J., and Sanderson, M. J. (2008). Penalized likelihood phylogenetic inference: Bridging the parsimony-likelihood gap. *Syst. Biol.* 57, 665–674. doi: 10.1080/10635150802422274

Konieczny, A., and Ausubel, F. M. (1993). A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. *Plant J.* 4, 403–410. doi: 10.1046/j.1365-313X.1993.04020403.x

Kreft, H., and Jetz, W. (2007). Global patterns and determinants of vascular plant diversity. *Proc. Natl. Acad. Sci. U. S. A.* 104, 5925–5930. doi: 10.1073/pnas.0608361104

Lagomarsino, L. P., Condamine, F. L., Antonelli, A., Mulch, A., and Davis, C. C. (2016). The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). *New Phytol.* 210, 1430–1442. doi: 10.1111/nph.13920

Landis, J. B., Soltis, D. E., Li, Z., Marx, H. E., Barker, M. S., Tank, D. C., et al. (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* 105, 348–363. doi: 10.1002/ajb2.1060

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., 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

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

Lemmon, E. M., and Lemmon, A. R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 44, 99–121. doi: 10.1146/annurev-ecolsys-110512-135822

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Maas, P. J. M. (1972). Costoideae (Zingiberaceae). *Flora Neotrop.* 8, 1–139.

Maas, P. J. M. (1977). Renealmia (Zingiberaceae-Zingiberoideae) Costoideae (Additions)(Zingiberaceae). *Flora Neotrop.* 18, 1–218.

Maas-van de Kamer, H., Maas, P. J. M., Wieringa, J. J., and Specht, C. D. (2016). Monograph of African Costaceae. *Blumea J. Plant Taxon. Plant Geogr.* 61, 280–318. doi: 10.3767/000651916X694445

Maddison, W. P., and Knowles, L. L. (2006). Inferring Phylogeny Despite Incomplete Lineage Sorting. *Syst. Biol.* 55, 21–30. doi: 10.1080/10635150500354928

Maddison, W. P. (1997). Gene Trees in Species Trees. *Syst. Biol.* 46, 523–536. doi: 10.1093/sysbio/46.3.523

Mai, U., and Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* 19, 272. doi: 10.1186/s12864-018-4620-2

Mangiafico, S. S. (2016). Summary and analysis of extension program evaluation in R. *Rutgers Coop. Ext. New Brunswick NJ U. S. A.* 125, 16–22.

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

Matzke, N. J. (2013). Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. *Front. Biogeogr.* 5, 242–248. doi: 10.21425/f5fbg19694

McCormack, J. E., and Faircloth, B. C. (2013). Next-generation phylogenetics takes root. *Mol. Ecol.* 22, 19–21. doi: 10.1111/mec.12050

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., and Yang, Y. (2018). Practical considerations for plant phylogenomics. *Appl. Plant Sci.* 6, 1–15. doi: 10.1002/aps3.1038

Meseguer, S. A., and Condamine, F. L. (2020). Ancient tropical extinctions at high latitudes contributed to the latitudinal diversity gradient. *Evolution* evo.13967. doi: 10.1111/evo.13967

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2011). "The CIPRES science gateway: A community resource for phylogenetic analyses," in *Proceedings of the TeraGrid 2011 Conference: Extreme Digital Discovery, (TG'11).* (New York, NY: Association for Computing Machinery), 41, 1–8. doi: 10.1145/2016741.2016785

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., and Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics* 24, 1757–1764. doi: 10.1093/bioinformatics/btn322

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

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

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., et al. (2001). Terrestrial Ecoregions of the World: A New Map of Life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *Bioscience* 51, 933–938. doi: 10.1641/0006-3568(2001)051[0933:teotwa]2.0.co;2

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

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

Portik, D. M., Smith, L. L., and Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Mol. Ecol. Resour.* 16, 1069–1083. doi: 10.1111/1755-0998.12541

Quintero, I., and Jetz, W. (2018). Global elevational diversity and diversification of birds. *Nature* 555, 246–250. doi: 10.1038/nature25794

R Core Team (2013). *R: A language and environment for statistical computing*. (Vienna, Austria: R Foundation for Statistical Computing).

Ree, R. H., and Sanmartín, I. (2018). Conceptual and statistical problems with the DEC+J model of founder-event speciation and its comparison with DEC via model selection. *J. Biogeogr.* 45, 741–749. doi: 10.1111/jbi.13173

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

Rosas-Guerrero, V., Aguilar, R., Martén-Rodríguez, S., Ashworth, L., Lopezaraiza-Mikel, M., Bastida, J. M., et al. (2014). A quantitative review of pollination syndromes: Do floral traits predict effective pollinators? *Ecol. Lett.* 17, 388–400. doi: 10.1111/ele.12224

Rowan, B. A., Seymour, D. K., Chae, E., Lundberg, D. S., and Weigel, D. (2017). "Methods for genotyping-by-sequencing," in *Methods in Molecular Biology*. Eds. J. White Stefan and S. Cantsilieris (New York, NY: Humana Press Inc.), 221–242. doi: 10.1007/978-1-4939-6442-0_16

Salzman, S., Driscoll, H. E., Renner, T., André, T., Shen, S., and Specht, C. D. (2015). Spiraling into History: A Molecular Phylogeny and Investigation of Biogeographic Origins and Floral Evolution for the Genus *Costus*. *Syst. Bot.* 40, 104–115. doi: 10.1600/036364415x686404

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, 1–19. doi: 10.1371/journal.pone.0043808

Sass, C., Iles, W. J. D., Barrett, C. F., Smith, S. Y., and Specht, C. D. (2016). Revisiting the Zingiberales: Using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* 2016, 1–17 doi: 10.7717/peerj.1584

Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics* 27, 592–593. doi: 10.1093/bioinformatics/btq706

Serrano-Serrano, M. L., Rolland, J., Clark, J. L., Salamin, N., and Perret, M. (2017). Hummingbird pollination and the diversification of angiosperms: An old and successful association in Gesneriaceae. *Proc. R. Soc B Biol. Sci.* 284, 1–10. doi: 10.1098/rspb.2016.2816

Smit, A. F. A., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0. Available at: http://www.repeatmasker.org (Accessed March 15, 2020).

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

Soto Gomez, M., Pokorny, L., Kantar, M. B., Forest, F., Leitch, I. J., Gravendeel, B., et al. (2019). A customized nuclear target enrichment approach for developing a phylogenomic baseline for Dioscorea yams (Dioscoreaceae). *Appl. Plant Sci.* 7, 1–13. doi: 10.1002/aps3.11254

Specht, C. D. (2006). Systematics and evolution of the tropical monocot family Costaceae (Zingiberales): a multiple dataset approach. *Syst. Bot.* 31, 89–106. doi: 10.1600/036364406775971840

Specht, C. D., Yockteng, R., Almeida, A. M., Kirchoff, B. K., and Kress, W. J. (2012). Homoplasy, Pollination, and Emerging Complexity During the Evolution of Floral Development in the Tropical Gingers (Zingiberales). *Bot. Rev.* 78, 440–462. doi: 10.1007/s12229-012-9111-6

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. (1970). Adaptive Radiation of Reproductive Characteristics in Angiosperms, I: Pollination Mechanisms. *Annu. Rev. Ecol. Syst.* 1, 307–326. doi: 10.1146/annurev.es.01.110170.001515

Streicher, J. W., and Wiens, J. J. (2017). Phylogenomic analyses of more than 4000 nuclear loci resolve the origin of snakes among lizard families. *Biol. Lett.* 13, 1–7. doi: 10.1098/rsbl.2017.0393

Tonini, J., Moore, A., Stern, D., Shcheglovitova, M., and Ortí, G. (2015). Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. *PloS Curr.* 7, 1–15. doi: 10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be

Tripp, E. A., and Manos, P. S. (2008). Is floral specialization an evolutionary dead-end? Pollination system transitions in *Ruellia* (Acanthaceae). *Evolution* 62, 1712–1737. doi: 10.1111/j.1558-5646.2008.00398.x

Valderrama, E., Richardson, J. E., Kidner, C. A., Madriñán, S., and Stone, G. N. (2018). Transcriptome mining for phylogenetic markers in a recently radiated genus of tropical plants (Renealmia L.f., Zingiberaceae). *Mol. Phylogenet. Evol.* 119, 13–24. doi: 10.1016/j.ympev.2017.10.001

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, 1–14. doi: 10.1002/aps3.1036

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

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

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bull.* 1, 80–83. doi: 10.2307/3001968

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628

Zerbino, D. R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhang, C., Sayyari, E., and Mirarab, S. (2017). "ASTRAL-III: Increased scalability and impacts of contracting low support branches," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Switzerland: Springer Verlag), 53–75. doi: 10.1007/978-3-319-67979-2_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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership